CR-0364-1

HISTORICAL SIMULATION: A PROCEDURE FOR THE EVALUATION OF ESTIMATING PROCEDURES

VOLUME I OF II

PROCEDURE DEVELOPMENT AND DESCRIPTION

Final Report
Contract DAHC15-68-C-0364

June 1969

C. A. Graver

GENERAL
RESEARCH CORPORATION
P O BOX 3587, SANTA BARBARA, CALIFORNIA 93105

Prepared for
Office of Assistant Secretary of Defense
(Systems Analysis)
Washington, D.C.

UNCLASSIFIED

CR-0364-1

HISTORICAL SIMULATION: A PROCEDURE FOR THE EVALUATION OF ESTIMATING PROCEDURES

Vol. I of II

PROCEDURE DEVELOPMENT AND DESCRIPTION

Final Report
Contract DAHC15-68-C-0364

June 1969

C. A. Graver

Prepared for
Office of Assistant Secretary of Defense
(Systems Analysis)
Washington, D.C.

UNCLASSIFIED

ABSTRACT

A recurring problem faced by many analysts is that of devising estimating procedures for predicting some aspect of the future from rather meager data. This is particularly true for the cost analyst who is concerned with estimating the resource requirements of future military systems.

Historical Simulation is a method of evaluating candidate (cost) estimating procedures on the basis of their ability to simulate predictions using data that would have been available. For example, assume that a particular data base consists of perhaps 15 data points ordered in time; a typical simulated prediction would entail using a candidate estimating procedure to predict point 10 using only the information available in the first nine data points. All candidate estimating procedures would then be evaluated on how well their simulated predictions compare with the actual data points.

In this fashion, Historical Simulation avoids relying on the central evaluation assumption of Regression Theory, namely, that which fits the past data best will predict the future best. This conceptual difference gives Historical Simulation several unique features, among which are

1. The demonstration of an estimating procedure's capability to make predictions of those points in the data base which are extrapolations from the previous data
2. The ability to directly compare a wider class of estimating procedures than can be compared by the usual regression techniques

Preceding page blank

3. The ability to evaluate estimating procedures derived from stepwise regression independent of the selection process utilized in that technique

4. The use of an easy-to-communicate summary statistic fcr describing the accuracy of predictions.

Hence, Historical Simulation provides additional information which, when used in conjunction with the usual regression techniques, should lead to a better evaluation of candidate estimating procedures, particularly when the prediction problem is characterized by extrapolation from a small data base.

The report is in two volumes. The first, which is unclassified, completely describes the technique. Included is a discussion of reasons leading up to the development of Historical Simulation as well as a description of the technique and of possible ways to summarize and interpret the output. Volume 2, classified Confidential (Privileged Information), illustrates the use of Historical Simulation by describing the results of applying the technique to cost and man-hour estimating procedures for selected aircraft programs.*

*The reader interested largely in a nontechnical overview may prefer C.A. Graver, Progress Report On The Development of Historical Simulation, General Research Corp IMR-950, March 1969, which was delivered at the 1969 DoD Cost Research Symposium.[1]

ACKNOWLEDGMENTS

To the best of my knowledge, the basic concept of Historical Simulation was first thought of by Mr. John Berterman of Planning Research Corporation and was later brought to my attention by Mr. Harry Piccariello of the Federal Office of Education. I wish to thank these two gentlemen for their initial efforts as well as their comments concerning my work.

I would also like to thank the many individuals who have taken the time to comment on the development of Historical Simulation. In particular, I wish to acknowledge the contributions of Dr. Edward Dodson of General Research Corporation and Dr. George Patton of the Office of the Assistant Secretary of Defense for Systems Analysis. In addition, I would like to thank Professor Ingram Olkin of the Stanford University Dept. of Statistics for reviewing the derivations of Appendix II and suggesting several of the areas for future work.

I would also like to thank my secretary, Miss Nancy Ray, for the many hours that she worked on the preparation of this report.

CONTENTS

Preceding page blank

CONTENTS (Cont.)

ILLUSTRATIONS

TABLES

TABLES (Cont.)

I. INTRODUCTION

The purpose of this report is to describe the progress made in the development of Historical Simulation, a procedure for the evaluation of Cost Estimating Procedures (or Cost Estimating Relationships). The work is being sponsored by the Director of Economics and Resource Analysis, Office of the Assistant Secretary of Defense (Systems Analysis) under Contract Number DAHC15-68-C-0364.

As parts of this report are fairly technical, the reader interested largely in a nontechnical overview of the Historical Simulation procedure is referred to the paper delivered at the recent 1969 DoD Cost Research Symposium (Ref. 1 of this volume).

A. BACKGROUND

The current emphasis on systems analysis, while it has greatly enhanced the decision-making capabilities of defense policy makers, has placed a difficult requirement on cost analysts. Working with functional cost models which utilize a description of the system in terms of its most basic physical or performance characteristics, the analyst is asked to make estimates which often require extrapolations from extremely meager data. These estimates are used in the evaluation of which candidate system is to be pursued.

Because the generally sparse nature of the data tends to obscure genuine functional trends, the analyst must go to great pains to fully utilize all the information his data base contains. While the cost analyst has at his disposal a number of tools, e.g., linear regression techniques, any additional tool that summarizes different information from the data base, such as Historical Simulation promises to be, is worthwhile.

Traditionally, in the development of a cost estimating relationship (CER), the cost analyst first postulates a functional relationship that hopefully will reflect the cost generating relationship underlying the data. The data base is then used to estimate the parameters of the functional relationship and a CER is obtained. Whenever the functional relationship is linear or can be transformed into a linear form, a least squares curve-fitting technique is generally used to estimate the parameters. At this point the analyst may examine any of a number of measures, or statistics, based on linear regression theory to assess the goodness of the resulting fit. If the fit is judged good the analyst uses the CER as the basis for cost prediction, concluding that it represents the cost generating process of the class of systems being analyzed. The assumption operating here is, in effect, that which fits the data best predicts best.

While no necessary relationship to the system's cost generating process is thus established, a good fitting CER can be meaningfully used to make cost predictions, particularly when the desired prediction is an interpolation within the framework of the data base. But cost analysts often deal in extrapolations. New systems are generally bigger, or faster, or newer in some combination of physical or performance characteristics, and so fall outside existing data. Hence, to predict the cost of a future procurement, the cost analyst is often required to extrapolate from the past data base.

Historical Simulation extracts information from the data base on how well a cost estimating procedure[*] has performed similar extrapolations. However, Historical Simulation cannot guarantee (any more than regression techniques can) that an apparently valid cost estimating procedure can predict accurately a future procurement, as the cost generating process underlying this procurement may have drastically changed from the one underlying similar objects already produced. What is unique about Historical Simulation is that in evaluating the candidate cost estimating

[*] The functional form together with the technique for picking the parameters, as distinct from a particular CER which is the functional form together with estimated parameters.

procedure it directly uses the cost analysts goal--predicting costs of future objects using past data on similar objects.

The premise underlying Historical Simulation lies in the observation that, if the hypothesized functional relationship represents the cost generating process, and if the parameter estimating technique is valid, then the estimating procedure's validity can be demonstrated by simulating predictions* that might have been made using it throughout the time period of the data base. The resulting predictions can then be compared with actual data. Thus an analyst can test his estimating procedure by using some of his data to simulate a prediction of a later data point. If such simulated predictions yield consistently acceptable predictions, his confidence in the estimating procedure's ability to predict future procurements is greatly bolstered, even if the future procurement lies outside the data base.

In contrast to the that-which-fits-best-predicts-best rationale of linear regression theory, the assumption implicit in the Historical Simulation approach is that which simulates its ability to predict best will continue to predict best. This conceptual difference will provide the five advantages listed below:

1. The past ability of candidate cost estimating procedures to extrapolate from historical data can be demonstrated.

2. Evaluations made using Historical Simulation constitute additional information useful in hypothesizing new cost estimating procedure candidates.

*The use of the word prediction in the Historical Simulation context, may or may not have the usual meaning. If the candidate cost estimating procedure is hypothesized independent of the data base, then the simulated predictions are in fact predictions. But in the most typical case, when the candidate cost estimating procedure is hypothesized after examining the entire data base, the simulated predictions cannot be interpreted as actual predictions, for the candidate estimating procedure undoubtedly fits the entire sample well.

3. Historical Simulation can compare a wider class of cost estimating procedures than the usually employed regression techniques.

4. CERs derived from stepwise multiple regression techniques can readily be tested, thus providing an independent evaluation of them.

5. Evaluations made using Historical Simulation yield an easy-to-communicate summary statistic that is useful in describing the accuracy of a prediction.

To summarize, the conceptual differences between Historical Simulation and Regression Theory insure that the former will give the cost analyst new information from which he can judge the reliability and validity of hypothesized cost estimating procedures. Hence Historical Simulation is not a replacement of the traditional Regression Theory techniques; rather it is another tool which the analyst can use.

B. ORGANIZATION OF THE REPORT

This report is presented in two volumes of which this is the first. The second, subtitled Some Examples, presents the results of applying the Historical Simulation technique to two aircraft samples. While the author is not sufficiently familiar with the data to draw concrete conclusions about which estimating procedure is best, the results are useful in demonstrating the value of Historical Simulation. Volume II carries a Confidential classification.

Volume I completely describes the Historical Simulation technique, and presents related background material about current estimating techniques. It is in five sections, of which this Introduction is the first, and has three appendixes.

Section II is in large measure devoted to background material, and outlines the considerations and problems that have led to the development of Historical Simulation. It is concluded by listing some of the properties that would be desired of any new evaluation procedure.

Section III describes the Historical Simulation procedure in detail, demonstrating its use with a hypothesized linear cost estimating procedure, and utilizing a least squares fitting technique to estimate the parameter values. It is then generalized to a wider class of estimating procedures and some of its properties are discussed.

Section IV discusses three of the ways the outputs provided by Historical Simulation can be utilized. These three ways, or categories, are (1) direct examination of the output, (2) data summarizations that do not depend on a particular cost estimating procedure, and (3) statistics which utilize the assumptions of a particular cost estimating procedure.

Section V concludes the body of Volume 1 with a discussion of the advantages and current limitations of Historical Simulation, and identifies some of the directions future research in the technique might take.

The three appendixes contain topics of special interest. Appendix I describes a computer model written for Historical Simulation; Appendix II derives the distribution of the Historical Simulation predictions and residuals under the usual regression theory assumptions; and Appendix III compares several variance estimators.

Before proceeding to the body of the report, it should be understood that the word simulation, as it is used here, refers to the demonstrating of a cost estimating procedure's predictive capability by simulating a prediction that might have been made using only the data that would have been available. Thus this procedure does not include generating a random sample needed for Monte Carlo evaluation--an integral feature of many simulation models.

In addition, while the present work is tailored to the cost problem, no limitation is evident that precludes using Historical Simulation to evaluate any estimating procedure, particularly when the inference to be made has the characteristics of extrapolation and small sample size.

II. BACKGROUND TO THE DEVELOPMENT OF HISTORICAL SIMULATION

A. ADVANTAGES OF FUNCTIONAL COST MODELS

In recent years, major procurement and force decisions in the Department of Defense have been made with the help of Systems Analysis, a management tool in which alternative weapon systems capable of accomplishing the same objective are compared analytically. The alternatives are most often described in terms of general performance characteristics. Thus a bomber might be described by its speed (Mach 1.2, say), range (1500 nautical miles), and payload (18,000 pounds).

Before the various alternatives can be compared, estimates of each system's cost and effectiveness must be made. From these estimates the "best" alternative can be selected or new alternatives specified and the process repeated.

Traditionally, cost estimates have been based on detailed engineering evaluations of the weapon system alternatives. Indeed this process is still used, particularly in industry, when the comparisons being made concern the detailed design decisions necessary to achieve the specified weapon system characteristics (in the most economical fashion). For example, What should be the shape of the wing?

However, for making cost estimates to be used in choosing the major performance characteristics of the weapon system best suited to a specific mission, it has been found that functional cost models have several advantages over the more traditional engineering approach. By including in a system's functional cost model all significant cost generating performance characteristics, the cost estimate will depend as much as possible upon the same variables used to generate the effectiveness estimates.

In addition, functional cost models provide the rapid estimating capability necessary for making timely comparisons between alternate weapon systems having widely varying performance characteristics. Cost estimates generated in this manner, when used in conjunction with effectiveness estimates, become an integral part of the weapon system performance characteristic specification, rather than remaining the result of a more detailed evaluation for a particular weapon system configuration (which has been chosen without regard to cost).

Finally, a functional cost estimating procedure guarantees a consistent evaluation of cost. This is not usually the case in engineering evaluations where cost definitions and accuracies used in the evaluation of a particular alternative may differ from those used in the study of another alternative.

B. CURVE FITTING AND REGRESSION TECHNIQUES IN FUNCTIONAL COST MODEL SPECIFICATION

At one time, only curve fitting techniques (such as least squares) were used to develop particular cost estimating relationships (CERs) in a functional cost model. But, by themselves, the fitting techniques would not tell the analyst anything about the reliability of cost estimates made using a particular CER, nor would they help him choose the best from several competing CERs.

Statistical regression techniques which essentially measure the goodness of fit were introduced to answer these questions. Statements concerning predictive reliability were derived by using R-scores and prediction intervals, while choices between CERs with different input variables but the same functional form (e.g., linear) were made using F- and t-tests.

There was, however, a certain amount of trial and error involved in applying these regression techniques. Candidate CERs had to be specified, and often the results of applying the regression techniques were such that none of the CERs were acceptable. A computer routine called Stepwise Regression[2] has been utilized by some to eliminate a great deal of the trial and error. The analyst has only to specify the candidate independent variables and desirable variable transformations (e.g., square root, squared, multiplication of two together, etc.) rather than to hypothesize the candidate CERs. The stepwise routine can evaluate various linear combinations of candidate variables and their transformations to derive one of the best* linear combinations (in the sense of fitting the data best) for a specified number of variables. The use of this program will be discussed further in Sec. II C 3.

C. PROPERTIES DESIRED IN ANY NEW EVALUATION PROCEDURE

The application of curve fitting and regression techniques has led to several problems, four of which are amenable to evaluation using Historical Simulation. The ability to deal with these problem areas is highly desirable in any new evaluation procedure; each is discussed below in terms of the stated requirements that any new evaluation procedure should have.

1. Needed: A Simple Measure to Define the Predictive Capability of Candidate Cost Estimating Procedures or CERs**

A problem in applying statistical regression techniques is that the cost analysis application is typically characterized by small sample sizes. Hence every attempt is made to build up the sample by including all data that is practically relevant. In so doing, however, the fulfillment of required assumptions, such as independence of sample observations, becomes

* There has been some discussion as to whether or not the resulting linear combination is the best. Step-forward and step-backwards routines do not always result in the same linear combination for K-variables. For further discussion see Ref. 3.

**The difference between these terms is given in the discussion of Property 2.

doubtful. For this and other reasons[*] the usual statistica. interpretation of the regression statistics (i.e., F- and t-tests, R-score) is open to question; statements about significance levels and prediction intervals may be meaningless.

Even when the cost application does not satisfy the regression theory assumptions, however, it is possible to use the regression theory machinery to devise measures that are free from a statistical interpretation and have a justifiable "geometric" interpretation. Such a geometrical interpretation is described in Ref. 4, pages 13-27. This interpreation has had little use since its presentation, probably because of its complexity and the lack of exact rules to be applied in its application.

If a simpler, heuristic measure can be defined, one which will enable the analyst to choose among alternative CERs and to say something about the reliability of the estimate, there will be no real advantage in striving for wide understanding of this geometrical interpretation. Such a measure, called Average Proportional Error, is identified and discussed in Sec. IV B of this report.

2. Needed: An Evaluation Procedure That Can Directly Compare a Broader Class of Candidate CERs (Called Cost Estimating Procedures)

Under (1) above the question of the meaning of the usual statistics in the cost analysis application was addressed. Here attention is focused on the comparability of these statistics. How does one choose between models of different functional form, e.g., $Y = a + bX$ and $Y = aX^b$?

One approach is to use the index of determination (or R^2). But the values of this statistic can not be directly compared and a model choice based on the index value closest to one can be very misleading.

[*] For a discussion of the regression theory assumptions and the question of whether they are satisfied in the cost analysis application, see Ref. 4, pages 3-8.

To illustrate this point, examine the contents of Table 1. This is the result of running a library computer program which fits six different curve forms (second column) in an attempt to choose the best. As can be seen by evaluating the indexes of determination (or R^2's--the column marked Index), curve six appears to be the best choice. A printout of the table of residuals quickly dispels this notion, however--the fit in terms of Y is lousy indeed.

The problem is that the indexes of determination are not comparable. This is because the index is calculated on a least squares fit. But the fit is not applied until the candidate curve has been transformed into a linear form. For example the linear form of Eq. 6 (Table 1) is

$$\frac{1}{Y} = A + \frac{B}{X} \tag{1}$$

The fit criterion then is

$$\sum_{i=1}^{n}\left(\frac{1}{Y_i} - A - \frac{B}{X_i}\right)^2$$

and A and B are picked to minimize this quantity. The index of determination is calculated on the linear fit and hence applies $1/Y$, and not to the quantity of interest Y. Hence, they should not be compared.*

*In fact, this particular example is a bad fitting technique, Examples of data that fit $Y = X/(AX + B)$ well, but do not fit Eq. 1 well, can be easily constructed.

The author is not asserting that valid comparisons for different functional forms cannot be made. For instance, in Ref. 5 valid comparisons are made for a linear model and an exponential model, i.e.,

$$Y = aX_1^{b_1} X_2^{b_2} \cdot \cdot \cdot X_p^{b_p} \qquad (2)$$

But these comparisons are based on either making the statistics comparable or making the parameter selection technique the same. In the example of Table 1, however, care is not taken to make the index of determination comparable even though the param er selection techniques are different; the curve-fitting technique is first a transformation of the equation, i.e., Y = X/AX + B becomes Eq. 1, and then a least squares curve fit.

TABLE 1

MODEL COMPARISONS

```
XMEAN:  7.5                                YMEAN:  114.79

NUMBER          CURVE              INDEX            A              B

 1              Y=A+B*X            .87Ø942       1.578Ø2        15.Ø134
 2              Y=A*EXP(B*X)       .734369       12.9491        .238688
 3              Y=A*X↑B            .943895       5.5841         1.46224
 4              Y=A+(B/X)          .644278       164.1Ø8       -214.977
 5              Y=1/(A+B*X)        .45971        .Ø93564       -8.48198 $-3
 6              Y=X/(A*X+B)        .982Ø73      -1.79869 $-2    .2Ø6394

FOR WHICH CURVE ARE DETAILS DESIRED (NUMBER) ? 6


COEFFICIENTS:

                EXPECTED VALUE        95PCT CONFIDENCE LIMITS

       A:      -1.79869  $-2     -2.38852 $-2     -1.2Ø886 $-2
       B:       .2Ø6394           .188814          .223974


X-ACTUAL         Y-ACTUAL           Y-ESTIM         95PCT CONFIDENCE LIMITS

 1               5.2                5.3Ø765        4.93682          5.73871
 2               11                 11.7357        1Ø.9222          12.68Ø1
 3               23.2               19.68Ø7        18.Ø428          21.6457
 4               44.1               29.7516        26.3993          34.Ø79
 5               76.4               42.9333        36.25            52.638
 6               116.4              6Ø.93Ø4        48.Ø256          83.3188
 7               141.3              86.9715        62.3615          143.668
 8               159.2              128.ØØ2        8Ø.2Ø55          336.774
 9               164.6              2Ø2.191        1Ø3.Ø34          5372.98
 10              167.8              376.996        133.284         -455.Ø22
 11              169                1288.27        175.282         -24Ø.812
 12              17Ø.4             -127Ø.Ø7        237.532         -172.871
 13              173.8             -473.845        339.343         -139.516
 14              176.1             -3Ø8.221        536.Ø1          -119.696
```

To further clarify this distinction it is helpful to make explicit the often-neglected difference between a CER and what I have called the Cost Estimating Procedure.

A cost estimating procedure consists of a parametric estimating relationship (PER) PLUS a technique for estimating the values of the parameters (in the PER) from some sample. Thus an example of an estimatir procedure might be:

Parametric Estimation Relationship: $Y = a + bX$

(Y is the production cost of the item to be estimated

X is the weight of the item to be estimated)

Technique: Least squares curve fit

A new estimating procedure results from choosing a new PER, a new technique, or both Hence the combinations given in Table 2 are all examples of alternative estimating procedures.

When a cost estimating procedure, with PER $Y = a + bX$, say, is used in conjunction with a particular sample, (i.e., a particular set of observations) there is .erived an explicit cost estimating relationship (CER), for example, $Y = 10 + 25X$. This is a result of estimating the PER parameters by applying the estimating technique to the given sample.

> Thus every CER has identified with it a particular sample and an estimating procedure consisting of a PER and a technique.

The relationship of these entities is pictured in Fig. 1.

The usual regression theory statistics are comparable if the technique is the same for all candidate estimating procedures. In particular, this is true if the candidates have the same PER form, as the same technique can easily be used. For example, one can compare a linear

TABLE 2

COST ESTIMATING PROCEDURES

Procedure Number	PER	Technique
1	$Y = a + bV$	Least squares fit
2*	$Y = a + bV$	Line determined by the closest two data points in terms of V
3*	$Y = a + bX$	Same as above except closest measured in terms of X
4	$Y = aX^b$	Least squares fit on $\log Y = \log a + b \log X$

Y = production cost, X = weight, V = volume

*Procedures 2 and 3 in Table 2 may need some explanation. The technique proposed is very close to costing by analogy. In effect, the analyst assumes that if he forms a line with the two closest data points (in terms of his independent variable) to the point he wishes to predict, the estimate using this line will be better than an estimate made using a line that fits all the data.

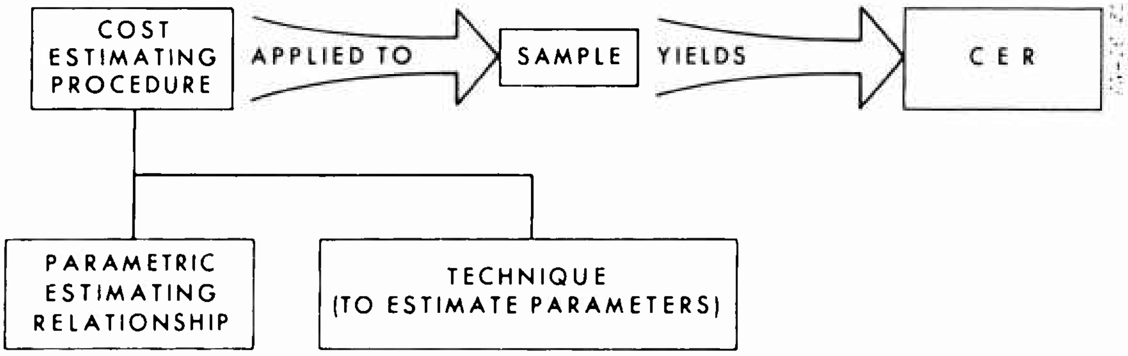
COST ESTIMATING PROCEDURE
APPLIED TO
SAMPLE
YIELDS
CER
PARAMETRIC ESTIMATING RELATIONSHIP
TECHNIQUE (TO ESTIMATE PARAMETERS)

Figure 1(U). Relationship of CER and Cost Estimating Procedure

PER which has two independent variables with one which has three independent variables.

If the PER forms are different, however, it is not always easy to choose the same technique. Applying least squares directly to a PER form such as Eq. 2 requires the use of expansions and iterative computer solutions.[5]

What is needed, then, is an evaluation procedure which can compare any cost estimating procedures without regard to whether or not the techniques are the same. As Sec. III D points out, Historical Simulation is such an evaluation procedure.

3. Needed: A Means of Evaluating Estimating Procedures Derived With Help of Stepwise Regression

Prior to the introduction of the stepwise regression technique, candidate CERs had to be hypothesized, with the hypotheses presumably based on engineering rationales or other criteria. The need for this specification was operationally removed when the stepwise multiple regression routine became available. Only the candidate variables and their allowable transformations had to be specified. However, when the stepwise routine was applied the resulting CER, while fitting the data well, often had no physical rationale. The applicability of the result then became questionable, even with a good fit. For example, suppose a hundred different CER combinations are tried. It is not surprising that one or two will fit well enough to be judged significant at the 0.05 significance level. This follows from the fact that the CER hypothesis is not picked *a priori* but is the result of finding the one that fits the data best from a hundred linear combinations; as such, this fit could easily represent one of the five times out of 100 that such a fit theoretically occurs by chance (at the 0.05 significance level).

With such misgivings concerning the results of stepwise regression, it would be valuable to have an evaluation procedure which could check estimating procedures derived by this technique. It will be shown in Sec. III D that Historical Simulation can make this independent evaluation.

4. Needed: An Evaluation Procedure Free From the *That-Which-Fits-Best-Predicts-Best* Curve-Fitting Assumption

The discussion above of the third desired property throws into doubt one of the central assumptions of least squares curve fitting--*that which fits the past best will predict the future best*--for it is this criterion that the stepwise regression procedure uses to choose CERs.

A second peculiarity of the cost analysis problem, in addition to small sample sizes, casts further doubt on the applicability of this least squares curve-fitting assumption. While using the criterion of *that which fits best, predicts best* should work reasonably well for cost predictions that are interpolations on the characteristics present in the data base, the criterion yields little information concerning cost predictions of procurements which represent extrapolations from the characteristics in the data base (see Ref. 6, page 6).

Predicting the cost of procurements that represent extrapolations from the data base is precisely the problem that the cost analyst usually faces. It seems like we are always required to estimate the cost of a bigger or faster plane, or one that is *better* in some combination of characteristics than those procured in the past.

Hence, a fourth desirable property for a new evaluation procedure is that it be independent of the assumption of *that which fits best predicts best*. In addition, it will be desirable that the evaluation procedure depends on how well the candidate cost estimating procedure can extrapolate from historical data. As will be seen in Sec. III D, Historical Simulation is such an evaluation procedure.

III. HISTORICAL SIMULATION DESCRIPTION

A. BASIC CONCEPT

The job of a cost analyst is to try to predict the cost (in constant dollars) of a proposed future procurement. He has at his disposal a description of the procurement in the form of a set of physical and performance characteristics. In addition, he has available physical and performance characteristics as well as cost data on similar past procurements.* Hence, his primary obje_tive is the prediction of a future procurement using available historical data.

Historical Simulation uses this primary objective in measuring the value of a cost estimating procedure. This basic tenet can be stated as follows:

> The cost estimating procedure which can best simulate predictions that would have been made in the past will actuall, be best able to predict the future.

B. AN EXAMPLE

To evaluate different cost estimating procedures, using the tenet just stated, Historical Simulation calls for each candidate cost estimating procedure to be tested on subsamples of the actual data base. For each subsample, the candidate cost estimating procedure is used to predict the cost of procurements built after any of the procurements in the subsample. These predictions are then compared to the actual costs.

* It will be assumed that the cost data is in constant dollars and pertains to some production quantity, like the hundredth unit.

To demonstrate this process, consider the following example comprising the thirteen data points listed in Table 3.[*] The data has been ordered as to date of first delivery (second column), and the actual cost and the independent variables X_1 and X_2 have been collected for each data point; X_1 and X_2 are physical or performance characteristics (such as weight and speed) which we hope will be useful in specifying the cost of the procurements we are to estimate. We have hypothesized the following cost estimating procedure:

$$\text{Cost} = a + b_1X_1 + b_2X_2 \tag{3}$$

where a, b_1 , and b_2 are to be estimated through the process of a least squares curve fit.

TABLE 3

SAMPLE DATA

Procurement Number	First Delivery	Actual Unit Cost	X_1	X_2
1	1950	95	1,996	153
2	1951	31	967	144
3	1953	60	2,414	149
4	1954	82	4,418	144
5	1956	25	852	107
6	1958	67	2,072	136
7	1960	243	10,408	177
8	1961	54	2,643	160
9	1962	112	3,786	172
10	1963	106	3,335	203
11	1964	183	6,374	196
12	1965	156	7,092	187
13	1967	177	10,304	167

[*] This data was used to debug the Historical Simulation computer program described in Appendix I. Values for Tables 3 through 8 were obtained from the output of this program as reproduced in Table 20 of Appendix I. The data used does not represent any real-world sample but is used only to illustrate the Historical Simulation procedure.

Now suppose we start with a subsample of five items; that is we will treat the first five rows of Table 3 as our data base. This is the data base from which an analyst would have had to make cost predictions in 1957. Using a least squares fit, the derived CER is

$$\text{Cost} = -73.9 + 0.0104X_1 + 0.792X_2 \qquad (4)$$

From Table 3, X_1 and X_2 for procurement number 6 are 2072 and 136. If these values are substituted into the CER of Eq. 4, the predicted cost is 55.3. From Table 3 the actual cost was 67; thus we have underestimated by 11.7.

Next, Eq. 4 can be used to predict the remaining data points 7-13. These predictions can be compared to the actual costs, and residuals calculated, yielding the results given in Table 4. As one can see there were six underestimates and two overestimates.

The entire process described thus far is now repeated for a subsample size of six. That is, we add the sixth procurement to our subsample, taking the six top rows of Table 3 as our data base. This data base is the one from which a cost analyst would have made his cost prediction in 1959. Making a least squares fit to this data base we obtain the following CER:

$$\text{Cost} = -68.4 + 0.0105X_1 + 0.765X_2 \qquad (5)$$

Comparing Eqs. (5) and (4) we see that the parameters have changed, although not by any great amount. This change is, of course, the result of adding procurement number 6 to the sample. The point to be remembered is that the explicit CER has changed, but the CER form, i.e., $\text{Cost} = a + b_1X_1 + b_2X_2$, and the parameter estimating technique, namely, least squares, has not changed. It is the CER form and the parameter estimating technique that are being evaluated by Historical Simulation, and not any one explicit CER such as Eq. 5.

TABLE 4

PREDICTED COSTS USING FIRST FIVE PROCUREMENTS

Procurement Number	Actual Unit Cost	Predicted Cost	Residual*
6	67	55.3	-11.7
7	243	174.4	-68.6
8	54	80.2	26.3
9	112	101.6	-10.4
10	106	121.5	15.5
11	183	147.5	-35.5
12	156	147.9	-8.1
13	177	165.4	-11.6

*Negative numbers are underestimates; positive numbers are overestimates.

Predictions and residual calculations for procurements 7-13 can now be made using Eq. 5 yielding the results shown in Table 5. Notice that procurement number 6 is not included since it was part of the data base used to derive Eq. 5.

TABLE 5

PREDICTED COSTS USING FIRST SIX PROCUREMENTS

Procurement Number	Actual Unit Cost	Predicted Cost	Residual
7	243	176.1	-66.9
8	54	81.7	27.7
9	112	102.8	-9.2
10	106	121.8	15.8
11	183	148.3	-34.7
12	156	149.0	-7.0
13	177	167.4	-9.6

The procedure described thus far can be repeated using subsample data base sizes of 7, 8, and on up to 13. In the last case the entire sample is usec and the usual least squares fit is obtained. Of course, no predictions for which an actual cost exists in the data base can be made using this final CER. However, this is the CER which will be used to make future predictions if the PER and parameter estimating technique being evaluated by Historical Simulation is chosen as a good method for predicting cost.

The outputs described can be conveniently summarized in a table of predictions (Table 6), a table of residuals (Table 7), and a table of parameter estimates (Table 8). The interpretation of this output will be discussed in Sec. IV.

A word of caution must be inserted at this point. The results of this particular example as displayed in Tables 6, 7, and 8 are merely illustrative. Their purpose is simply to make explicit the Historical Simulation procedure and the form of the output. Results of a limited number of Historical Simulation runs (using the computer program described in Appendix I) are presented in Volume 2 (CONFIDENTIAL) for some airciaft data. They were excluded from the present volume to avoid the necessity of classifying it.

Some of the possible ways of analyzing these results are discussed in Sec. IV, but it must be remembered that Historical Simulation is intended primarily as a tool for evaluating an estimating procedure. Hopefully, the evaluation will be made in the presence of other candidates. Only the analyst who understands his data base can make such judgements as to whether

- The results are reasonable, and the estimating procedure is valid, or
- The results are not reasonable and a new estimating procedure should be hypothesized, and/or the sample should be stratified --i.e., divided into groups which seem to come from different populations.

TABLE 6

PREDICTIONS

For Sample Point Number

Sample Size Used	6	7	8	9	10	11	12	13
5	55.3	174.4	80.2	101.6	121.5	147.5	147.9	165.4
6		176.1	81.7	102.8	121.8	148.3	149.0	167.4
7			85.4	114.3	128.1	177.4	183.9	227.0
8				102.1	103.9	161.5	172.6	229.3
9					110.7	166.3	176.1	229.2
10						164.5	174.9	229.8
11							179.7	223.7
12								227.1
13								

TABLE 7

RESIDUALS

For Sample Point Number

Sample Size Used	6	7	8	9	10	11	12	13
5	-11.7	-68.6	26.3	-10.4	15.5	-35.5	-8.1	-11.6
6		-66.9	27.7	-9.2	15.8	-34.7	-7.0	-9.6
7			31.4	2.3	22.1	-5.6	27.9	50.0
8				-9.9	-2.1	-21.5	16.6	52.3
9					4.7	-16.7	20.1	52.2
10						-18.5	18.9	52.8
11							23.7	56.7
12								50.1
13								

TABLE 8

ESTIMATED PARAMETERS

Sample Size	a	b_1	b_2
5	-73.9	0.0104	0.792
6	-68.4	0.0105	0.765
7	-74.6	0.0178	0.706
8	-31.8	0.0198	0.344
9	-45.2	0.0193	0.450
10	-38.6	0.0196	0.400
11	-50.2	0.0198	0.478
12	-44.6	0.0191	0.448
13	-63.9	0.0159	0.629

C. SUMMARIZATION OF THE PROCEDURE

This summarization, or generalization of Historical Simulation is presented in the language of the estimating procedures introduced in Sec. II in order to make it apparent that Historical Simulation can be used on any estimating procedure. (This was the second desirable property stated in Sec. II.)

Let the estimating procedure being examined have a PER given by

$$y = f(\vec{\beta}, \vec{X}) \tag{6}$$

where y is the cost

$\vec{\beta}$ are the parameters of the function

and $\vec{X}$ are independent variables

In the example given in Sec. III B, $\vec{\beta}$ represents the parameters a, b_1, and b_2 ; $\vec{X}$ the independent variables X_1 and X_2 ; and f the linear equation given by Eq. 3. To complete the estimating procedure specification, there is a technique T which, when applied to a sample, yields an estimate of the parameters $\hat{\vec{\beta}}$. In the example of Sec. III B the technique T was least squares curve fit.

The sample consists of N sets of data $(y_i, \vec{X}_i)$, $i = 1, 2, \ldots, N$, where the y_i are the actual cost of procurement i , and the $\vec{X}_i$ are the values of the independent variables for procurement i . It is assumed that the sample has been ordered in time, with the smaller values of i corresponding to the older data points.

The Historical Simulation procedure can then be summarized as an iterative process which goes through the following four steps at each iteration.

Step 1. Subsample Specification: Determine data base size n for this iteration, where n is larger than the subsample size of the previous iteration. In particular $n_o \leq n < N$ where n_o is some minimum sample size which is greater than the number of PER parameters, i.e., entries in $\vec{\beta}$. In the case of the example, $n_o \geq 4$ as there are three parameters to estimate: a, b_1 , and b_2.

Step 2. CER Specification: Apply the estimating procedure technique T to the subsample of size n identified in Step 1, i.e., $(y_i, \vec{X}_i)$; $i = 1, 2, \ldots, n$, and obtain the PER parameter estimates $\hat{\vec{\beta}}_n$. (In the example of the last section, least squares estimates of a, b_1 , and b_2 were made for each iteration.) Substituting these parameter estimates into the PER yields the CER for this iteration. It can be denoted by

$$\hat{Y} = f\left(\hat{\vec{\beta}}_n, \vec{X}\right)$$

Step 3. Cost Prediction: Predict the cost of each of the procurements not included in the subsample. This is accomplished by substituting the values of the independent variables (for the procurement in question) into the CER developed in Step 2. Predictions are made of y_{n+k}, $k = 1, 2, \ldots, N-n$. These predictions are labeled $\hat{y}^{(n)}_{n+k}$ in the remainder of this report and are given by

$$\hat{v}^{(n)}_{n+k} = f\left(\vec{\beta}_n, \vec{X}_{n+k}\right) \quad ; \quad k = 1, 2, \ldots, N-n \tag{7}$$

where $\hat{y}^{(n)}_{n+k}$ is the prediction of y_{n+k} from subsample size n and the $\vec{X}_{n+k}$ are the values of the independent variables for the n+kth procurement. (For the example these predictions were listed in Table 6.)

Step 4. Calculation of the Residuals: The actual costs are subtracted from the appropriate predictions (Step 3) and the residuals obtained. These residuals, denoted by $d^{(n)}_{n+k}$, are given by

$$d^{(n)}_{n+k} = \hat{y}^{(n)}_{n+k} - y_{n+k} \tag{8}$$

Negative values of $d^{(n)}_{n+k}$ represent underestimates while positive values are overestimates. (The residuals for the example of the last section were given in Table 7.)

A few remarks should be made concerning Step 1, Subsample Specification. For the purposes of Historical Simulation, several data points procured in the same time frame can be grouped together.* For instance, if data points--procurements--7, 8, and 9 were all delivered in the same year, one can group this data. Iterations of the Historical Simulation

*Grouping will have no effect on the Historical Simulation evaluation with the exception of those statistics discussed in Sec. IV C 2 which, at present, are valid only for the one-step residuals $d^{(n)}_{n+1}$.

would include subsamples 5, 6, 9, 10, 11, 12, and 13. Predictions of data points 7, 8, and 9 would only be made with subsamples of five and six data points. Information concerning data points 7 and 8 would not have been available for the prediction of data point 9, so grouping the data does not invalidate the Historical Simulation procedure.

Another problem in subsample specification is selecting the initial subsample. A lower bound exists that is dictated by the number of parameters to be estimated. For the example in this section the lower bound would be four (one greater than the number of parameters as required for a finite variance least squares fit). But this selection of four subsample items would allow only one degree of freedom and one would expect a great deal of variation in the predictions. Using too large an intial sample, however, will greatly reduce the amount of new information contained in Tables 6, 7, and 8. The initial subsample size must thus be set by the analyst at the smallest number which is necessary for the estimating procedure, if valid, to have enough information from which to make reasonable estimates. (In the example n_o was arbitrarily chosen to be 5).

D. SOME PROPERTIES

Several properties of the Historical Simulation procedure can be established from the development made thus far. For instance, the procedure evaluates a candidate cost estimating procedure by simulating how well the latter would have predicted if it had been available and used to make cost estimates in the past. Hence the name Historical Simulation.

Historical Simulation does not depend on the usual curve fitting assumption of goodness *that which fits best, predicts best.* (This was identified as desirable property number 4 for a new evaluation procedure in Sec. II C). The freedom from the curve fitting assumption is a consequence of the fact that the output in Tables 6 and 7 depends only on how well the hypothesized cost estimating procedure predicts. The entries do not depend on how well a particular CER fits the subsample that was given to it.

Historical Simulation can be used to evaluate CERs derived with the help of Stepwise Multiple Regression programs, whose use was discussed briefly in Sec. II. Using this program, the choice of a CER is determined by which candidate CER fits the data best (in a least squares sense). Unfortunately, the values of the usual regression statistics depend on this choice criterion and are thus not independent of the CER selection process. In contrast, Historical Simulation does not depend on the choice criterion as its output does not depend on how well the CER fits. In other words, Historical Simulation, unlike the usual regression statistics, is able to evaluate the CER independently of the stepwise regression choice criterion. (This property was identified as desirable property number 3 for a new evaluation procedure in Sec. II C.)

Due to its dependence on predicting from past data, Historical Simulation is a tool to demonstrate the estimating procedure's ability to handle extrapolations implicit in the data base. This is in contrast to the estimating procedure's ability to interpolate, which can be evaluated by the usual regression theory approach. The extrapolation is in the time direction as the data is ordered on time. Indeed this is probably the most universal ordering as it will tend to parallel orderings on physical characteristics. This is because new procurements usually represent advancements in the state of the art, as measured by some set of physical characteristics. Hence ordering on time will also tend to order on these physical characteristics.*

*It should be noted that there may be applications in which the advancement implicit in a new procurement is represented by an increase in one physical characteristic, say bandwidth. The problem then would be to estimate the cost of this new procurement, from a data base of procurements which all have smaller bandwidths. The extrapolation then would be in the bandwidth direction and, in this case, the author sees no reason why the ordering could not be on bandwidths.

Another difference between Historical Simulation and the usual curve fitting techniques is that the former looks at different samples while the latter concentrates on the entire historical sample. In effect the Historical Simulation procedure looks at how well the hypothesized CER form does at varying times and hence how reliable the hypothesized CER is over time. In contrast, the curve fitting techniques and the associated regression statistics evaluate one period in time, the present, and will in general* be unable to detect time-trend effects.

Finally, Historical Simulation can be used to directly compare any candidate cost estimating procedures. (Identified in Sec. II C as desirable property (2) for a new evaluation procedure.) This is quite apparent from the fact that the summarization of the procedure in the last section was carried out in estimating procedure notation. All that is needed is a PER, Eq. 7, and a parameter estimating technique T.

Having defined the Historical Simulation procedure and some of its properties and seen how it works for a particular example, attention must now be focused on the output of Historical Simulation. What is it good for and how does one interpret it? These questions will be addressed in the following section.

*This statement is not universal because time has sometimes been included explicitly in the CER form.

IV. OUTPUT INTERPRETATION

In trying to interpret the results of Historical Simulation (or indeed to make inferences from the usual regression statistics), the analyst is trying to examine two basic questions about the cost estimating procedure under study:

1. Is the estimating procedure valid?, i.e., is it a true representation of the cost generating process under study?
2. How reliable is the estimating procedure?, i.e., is the model variance, and hence the variance in estimates, large or small?

Insights into the answers to these questions are used by the analyst to choose between different candidate cost estimating procedures (ranking), to define new candidate cost estimating procedures, and to make statements about the accuracy of his predictions.

The value of the Historical Simulation procedure must be directly related to the usefulness of its output as a means of providing insights into these two basic questions and helping the analyst make the choices and statements identified above. Ways of using the Historical Simulation output for these purposes are discussed in this section. The discussion has been organized into the following three categories:

1. Direct Examination of the Historical Simulation Output (Sec. IV A)
2. Data Summarizations That do not Depend on a Particular Estimating Procedure (Sec. IV B)
3. Statistics Which Depend on a Particular Estimating Procedure (Sec. IV C)

A. DIRECT EXAMINATION OF THE HISTORICAL SIMULATION OUTPUT

A direct examination of the contents of the output tables of Sec. III (Tables 6, 7 and 8), can add insight into the question of model validity, the identification of questionable sample points, and the identification of new candidate estimating procedures. In the course of this examination several useful questions can be asked; these are discussed below making use of the form of the residual table (Table 9) which is patterned after Table 7 of Sec. III.

TABLE 9

FORM OF RESIDUAL TABLE

(X stands for a residual value calculation)

Sample Sizes Used	Sample Point							
	6	7	8	9	10	11	12	13
5	X	X	X	X	X	X	X	X
6		X	X	X	X	X	X	X
7			X	X	X	X	X	X
8				X	X	X	X	X
9					X	X	X	X
10						X	X	X
11							X	X
12								X
13								

1. Each column of Table 9 gives the residuals for a particular sample point. One can ask if these residuals are <u>improving</u> --getting smaller in an absolute sense--as the sample size grows (that is, as the analyst looks down the column). One would expect the residuals to improve--or at least not get any worse--if the model is valid and the sample consistent.

   In Table 7 we saw that this behavior is not true for the test run sample. The residuals are erratic or tend to get worse for sample points 9, 11, 12 and 13.

2. Are there any consistent errors? For example, does the estimating procedure underestimate (have negative residuals) most sample points consistently? If so, then the cost

estimating procedure shows signs of bias. Again by examining any column of Table 9, one might find sample points that are consistently under- or over-estimated by a substantial amount. In this case there is reason to suspect that the data point in question does not belong to the population, or that errors have been made in recording its cost or the values of the independent variables.

For the test run data of Table 7 there appears to be no indication of bias as the residuals are neither mostly negative or mostly positive. There are sample points, however, that show substantial consistent errors, such as points 7 and 11.

3. Residuals along any row of Table 9 are all derived from the same subsample. Comparing two adjacent rows indicates the impact on the prediction process of the points added to the larger subsample. One might therefore ask if there have been significant changes, in some consistent manner, from one row to the next. If so, the sample point added is dominating the estimating procedure and if the changes in residuals are not for the better (i.e., smaller absolute residuals) then the question of whether or not the sample point properly belongs to the population is again raised.

As an example, if rows for subsample sizes of six and seven data points are compared in Table 7, we see substantial changes in the residuals. While some residuals have improved --sample points 9 and 11--others have definitely become worse --sample points 12 and 13. There is no question that sample point 7 has had a significant impact, but its impact is mixed.

4. Finally, the estimates of the parameters (Table 8) can be examined. Are they reasonably stable, showing signs of convergence as the sample size grows? If so, then one feels a greater assurance of the model's validity; the information concerning the values of the model parameters is essentially

the same from all the sample points. If not, then there might be something in the pattern of the estimated coefficients that would suggest a new candidate cost estimating procedure or that would identify a questionable sample point.

In Table 8 it can be seen that the desired stability did not take place for the test run data. The inclusion of sample points 7 and 8 had a significant impact on the parameter estimates to the subsample 7 values. Hence, these points ough to be examined carefully.

In summary, there is a great deal of "look-see" evidence concerning the model validity in the output of Historical Simulation. This output can be used to build confidence in model validity or, conversely, aid in hypothesizing a new cost estimating procedure. In addition, it can help to identify questionable sample points. Furthermore, no information concerning the process has been lost. This is in contrast to the statistics discussed under the remaining two groupings which depend on summarizations of the data--and most data summarizations imply a loss of some information.

B. DATA SUMMARIZATIONS THAT DO NOT DEPEND ON A PARTICULAR ESTIMATING PROCEDURE

Data summarizations (or statistics) discussed in this section have the property that they can be calculated for any candidate cost estimating procedure. These summarizations can thus be used to compare different candidate estimating procedures.

This lack of dependence, however, introduces uncertainty as to what data summarizations should be used. The criteria required for measure selection, and the theoretical framework necessary for the description of measure properties, are usually provided by the form of the particular estimating procedure and the assumption of an underlying statistical model.

As an example, Multiple Linear Regression theory is based on a statistical model (assumptions) applicable to linear PERs. Using this model as a starting point, statistical arguments can be developed to pick the fit technique (least squares), to provide convenient summary statistics (t-tests, standard error of estimate, etc.), and to describe summary statistic properties.

Lacking the capability of specifying one "best" data summarization, several different summarizations are suggested in this section. Arguments for their use are necessarily heuristic in nature, and the choice of which particular summarization to use is left up to the analyst. He can exercise this choice by picking loss functions and weighting schemes best suited to his application.

Before describing the summarizations it will be useful to identify the portion of the Historical Simulation output that will be used. Only the values from the residual table--the $d_{n+k}^{(n)}$ of Eq. 8--are used as it is the errors of prediction that are of interest. Which of these residuals to use is not entirely clear.

Using all of the residuals is appealing in that no information will be thrown away. However, there are problems involved in knowing how to use all of them fairly. The residuals are certainly not independent, a fact that is proven under the usual regression assumptions in Appendix II. Hence, use of all of the residuals introduces problems of statistical interpretation and weighting.

If, however, only one residual is used for each sample point, in particular the one made from the largest available subsample size--the entry in the last column of Table 9, which is $d_{n+1}^{(n)}$ if there is no grouped data--then the problems of weighting and statistical interpretation* are

*In fact, it is shown in Appendix II that the one-step residuals $d_{n+1}^{(n)}$ are independent under the usual regression assumptions.

greatly reduced. Furthermore, this selection is not without heuristic justification. In effect, we are looking at the prediction made from the largest available subsample size for each procurement. These are the subsamples that would have been used and predictions that would have been made if the cost estimating procedure had been used in the past. In addition, an estimating procedure which predicts the near future well, need not necessarily predict the long term future well.*

For notational convenience, let us relabel these residuals by R_{n_o+1} , ... , R_n , ... , R_N , where n_o was the minimum sample size used in the Historical Simulation, and N is the size of the entire data base. The collection of these residuals will be referred to as $\vec{R}$.

The question being addressed in this section then is how to summarize the data in $\vec{R}$, so that one can choose between several estimating procedures. In addition, it will be useful if these summarizations indicate how well the estimating procedure will do in the future.

1. Some Example Data Summarizations

One such summarization is that of average proportional error. It is calculated as follows.

$$\text{Average Proportional Error} = \frac{1}{N-n_o} \sum_{i=n_o+1}^{N} \frac{|R_i|}{y_i} \qquad (9)$$

*This does not imply that all the predictions of the most recent data points are ignored. On the contrary, as will be seen in this section, predictions of the most recent data points will receive at least as much emphasis as predictions of the earlier data points. But the particular prediction used will be from the largest data base possible for such a prediction.

where y_i is the actual cost of the procurement indexed by i

n_o is the minimum sample size used in the Historical Simulation

and N is the size of the data base

The average proportional error should be used when one is worried about proportional cost errors rather than absolute cost errors. In addition, this measure is probably the easiest to communicate (and as such is a good candidate for the desired measure described in Sec. II C). Every cost analyst has been asked to indicate how reliable his prediction is; for example, is it within ±10 percent? Having calculated the average proportional error, he can answer this query by saying, "The cost estimating procedure from which this estimate has been derived has an average proportional error of, say 15 percent, which implies that if it had been used to make these types of predictions in the past it would have been off, on the average, by 15 percent." Hence, a reasonable answer to the query would be that an error of ±15 percent should be expected.

Contrast the above answer to one made from the usual regression theory output utilizing statements of F-tests, t-tests, R^2, prediction intervals, etc. How aware of the underlying statistical assumptions* or the meaning of these statistics is the recipient of the prediction results? Their meaning is certainly not as universally understandable as is average proportional error.

There are, of course, drawbacks in using averages associated with average proportional error, a topic which will be discussed more fully in Sec. IV B 3, Additional Considerations. In addition to these problems, however, average proportional error places the same emphasis on predictions made from a sample of size 5 as predictions made from a sample of size 12.

*See Ref. 4 for the interpretation of these statistics in cost analysis.

For any cost estimating procedure that makes use of every data point in its subsample, this equality of weighting may seem unjustified. After all, predictions should be getting better as the sample size increases. Hence, the following weighted average proportional error is suggested:

$$\text{Weighted Average Proportional Error} = \sum_{i=n_o+1}^{N} \frac{W_i |R_i|}{y_i} \qquad (10)$$

The weights, of course, add up to one

$$\sum_{i=n_o+1}^{N} W_i = 1$$

and varies proportionally with the sample size. They can be as extreme as assigning all weight to N , which is a choice that might be made by an analyst who feels that most information is contained in the one prediction made from the largest subsample size. My own preference for a weighting scheme is

$$W_i = \frac{S_i}{\sum_{i=n_o+1}^{N} S_i} \qquad (11)$$

where S_i is the subsample size used for the particular prediction. This equation would give the predictions from subsample size 10 twice as much weight as the predictions from subsample size 5, and thus is in accordance with the notion that if the estimating procedure is valid, then predictions should improve as the sample size gets larger. Furthermore, the use of this type of weighting scheme does not effectively change the simple interpretation of the summary statistic discussed for Eq. 9.

Another alternative to average proportional error is that of squared average proportional error, i.e.,

$$\text{Squared Average Proportional Error} = \frac{1}{N-n_o} \sum_{i=n_o+1}^{N} (R_i/y_i)^2 \qquad (12)$$

One would use this type of summarization when he wishes to penalize proportional errors in an exponential fashion.

Finally, one might be more concerned with absolute rather than relative error. A calculation such as

$$\text{Average Squared Error} = \frac{1}{N-n_o} \sum_{i=n_o+1}^{N} (R_i)^2 \qquad (13)$$

could be made. Although this statistic appears to be similar to the calculation of the variance estimate* in regression theory, the residuals in question here are based on predictions, not fits.

2. A General Framework for the Data Summarization

The data summarizations suggested so far can be placed into a general framework through the use of loss functions and weighting schemes. Let $\ell(R_i)$ denote the loss (or penalty) that will be assigned to the residual value R_i , and let W_i be the weight assigned to each residual, e.g., Eq. 11. Then the average loss for the weighting scheme W and the loss function ℓ can be defined by

$$A(\ell,W) = \sum_{i=n_o+1}^{N} W_i \ell(R_i) \qquad (14)$$

*Standard error of the estimate squared.

If the weight W_i could be interpreted as the probability of R_i occurring, then the average loss calculation, defined in Eq. 14, is equivalent to the calculation of expected loss in statistical decision theory (Ref. 7, Chapter 5). In this latter context, the decision rule (estimating procedure) with the smallest expected loss would be chosen. The analogous rule in the Historical Simulation context is to prefer the estimating procedure with the lowest average loss.*

Each of the example data summarizations previously specified is a special case of the generalized average loss identified in Eq. 14. Weighted Average Proportional Error, Eq. 10, is obtained by letting $\ell(R_i) = |R_i|/y_i$, while average proportional error, Eq. 9 implicitly uses the weighting scheme defined by

$$W_i = 1/N-n_o \tag{15}$$

This latter weighting scheme is used for each of the other averages previously discussed with the loss function defined by

$$\ell(R_i) = (R_i/y_i)^2 \quad \text{for Eq. 12}$$

and by

$$\ell(R_i) = R_i^2 \quad \text{for Eq. 13}$$

Any average loss can be used for ranking alternative estimating procedures. The analyst need only specify the loss function and weighting scheme best suited for his particular problem. For example, alternative loss functions might be devised to give a greater penalty to underestimates than overestimates. (All of the example loss functions previously

*Additional considerations are identified in Sec. IV B 3.

identified give equal penalty to these errors). Such a loss function is portrayed in Fig. 2 and defined by

$$\ell(R) = \begin{cases} R & \text{if } R > 0 \\ R^2 & \text{if } R \leq 0 \end{cases}$$

Furthermore the loss function need not be smooth. If one is very concerned about underestimates, doesn't care about overestimates with residual values of 0 to 15, and is only mildly concerned about greater overestimates, then the loss function given by

$$\ell(R) = \begin{cases} R-15 & \text{if } R \geq 15 \\ 0 & \text{if } 0 \leq R < 15 \\ R^2 & \text{if } R < 0 \end{cases}$$

could be used. This positive side of this loss function is shown as the dashed lines in Fig. 2.

There are some properties of specific weight and loss functions which in the author's mind make certain choices more natural than others. These considerations may help the analyst to choose the weighting scheme and loss function best suited for his application.

Regarding the weighting scheme, if the candidate estimating procedure makes use of the entire subsample, then the weights given in Eq. 11 appear most natural. It implies that the estimating procedure's predictive capability is directly proportional to the sample size.

UNCLASSIFIED

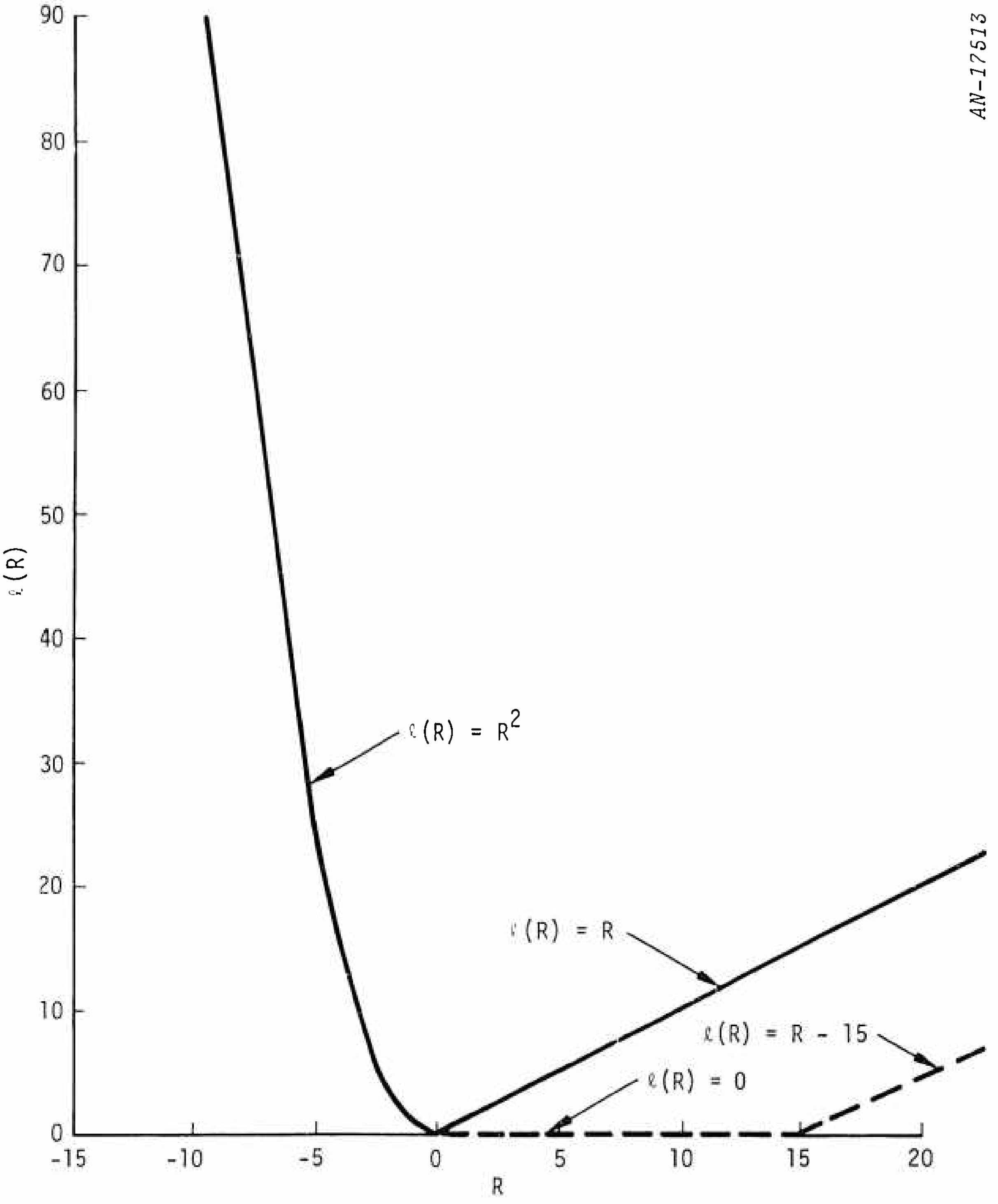

Figure 2(U). Greater Underestimate Penalty Loss Function

UNCLASSIFIED

A slight variation to this weighting scheme, but having similar properties, is one that is based on degrees of freedom. Let k be the number of parameters to be estimated in the candidate PER. Define W_i by

$$W_i = \frac{S_i - k}{\sum\limits_{i=n_o+1}^{N} (S_i - k)} \tag{16}$$

The weights are all positive since the minimum sample size n_o for Historical Simulation was defined in such a manner that $n_o > k$. Hence, $S_{n_o+1} > k$. This particular weighting scheme is analogous to adjusting for degrees of freedom in the usual regression statistics. It implies that the estimating procedure's predicitve capability is directly proportional to degrees of freedom.

A candidate es..mating procedure that does not make full use of the subsample at each stage of the Historical Simulation requires a different weighting scheme. For example, if the estimating procedure only makes use of the most recent four data points in each subsample, then a weighting scheme such as Eq. 15 would seem reasonable.

Regarding what loss function to select, if one is interested in relative error, then the loss function used might be that used in Average Proportional Error, Eq. 10 namely $\ell(R_i) = |R_i|/y_i$. It has the advantage of being easy to communicate as discussed in the paragraphs following Eq. 9.

If one is interested in absolute error, then the loss function used in Eq. 13, namely $\ell(R_i) = (R_i)^2$ could be used. It has the advantage of being the analogous calculation to the square of the standard error of the estimate from regression theory. The latter is the quantity minimized in least squares (if it were unadjusted for degrees of freedom) and hence has the advantage of precedent.

A disadvantage to this loss function is that it is not as easy to communicate as average proportional error. However the closely related loss function

$$\ell(R_i) = |R_i|$$

with average loss defined by

$$\text{Average Absolute Error} = \sum_{i=n_o+1}^{N} W_i |R_i| \qquad (17)$$

has the same meaning for absolute error as Eq. 10 has for relative error. It represents how much one would have been off (in an absolute sense) on the average, if he had used this cost estimating procedure consistently in the past.

The recommendations for loss functions and weighting schemes discussed in this section are summarized in Table 10. The reader is reminded that all of these summarizations are averages, and hence the decision rule of ranking the candidate estimating procedures and taking the one with the smallest average loss is an oversimplification of the problem, particularly when average losses are very close with the result that the difference may not be significant. Some further considerations that will help in the estimating procedure selection when the difference in average losses are small are given in the next subsection.

3. Additional Considerations

Suppose for a particular application average proportional error as given in Eq. 9 has been selected for the average loss calculation. Suppose also that estimating procedure A had an average proportional

TABLE 10

RESIDUAL SUMMARIZATIONS NOT DEPENDENT ON ESTIMATING PROCEDURE

$$\text{Average Loss} = \sum_{i=n_o+1}^{N} W_i \ell(R_i)$$

	Form	Remarks	
Suggested Weights	$W_i = \dfrac{S_i}{\sum_{i=n_o+1}^{N} S_i}$	Appropriate for estimating procedures which utilize entire subsample	Predictive capability of estimating procedure directly proportional to sample size
	$W_i = \dfrac{S_i - k}{\sum_{i=n_o+1}^{N} (S_i - k)}$		Predictive capability of estimating procedure directly proportional to degrees of freedom
	$W_i = \dfrac{1}{N-n_o}$	Appropriate for estimating procedures which utilize only the last m (any fixed number $\leq n_o$) subsample data points	
Suggested Loss Functions	$\ell(R_i) = \lvert R_i \rvert / y_i$	Appropriate for applications in which relative error is most important. Represents the average proportional error that we would have experienced if we had used the estimating procedure in the past.	
	$\ell(R_i) = R_i^2$	Appropriate for applications in which absolute error is most important	Analogous to the residual calculation in ordinary regression theory
	$\ell(R_i) = \lvert R_i \rvert$		Represents the average absolute error that we would have been off if the estimating procedure were used in the past

NOTATION

N Total data base size

n_o Minimum subsample size for Historical Simulation

R_i Residual, member of $\vec{R}$

S_i Subsample size used for R_i calculation

k Number of parameters estimated in the candidat PER

y_i Actual cost of procurement i

error of 0.2 and estimating procedure B had one of 0.25. Should A always be preferred to B ? At least two additional questions are worth asking.

1. Is there any apparent bias in the residuals?
2. What type of variability is there about the average loss?

The first of these considerations can be handled by a different type of average value calculation. Comparing the simple arithmetic mean of the residuals to zero could be used to indicate bias, if this calculation did not imply a weighting scheme and loss function different from the one picked by the analyst for the average loss calculation. Hence, to examine bias for our purposes, it is suggested that the following calculation be made.

$$B(\ell,W) = \sum_{i=n_0+1}^{N} W_i \ell^+(R_i) \qquad (18)$$

where W_i is the weighting scheme used in the average loss calculation of Eq. 14

ℓ^+ is a signed form of the loss function used in the average loss calculation

and $B(\ell,W)$ is the apparent bias of the estimating procedure using loss function ℓ and weighting scheme W.

Some explanation of $\ell^+(R_i)$ will be useful. If the average proportional error loss function is ℓ, i.e., $\ell(R_i) = |R_i|/y_i$, then $\ell^+(R_i) = R_i/y_i$. Hence the only difference between the two is that ℓ^+ retains the sign of the residual. For the squared error loss function $\ell(R_i) = R_i^2$ let

$$\ell^+(R_i) = \begin{cases} \ell(R_i) & \text{if } R_i \geq 0 \\ -\ell(R_i) & \text{if } R_i < 0 \end{cases}$$

Note that for any loss function, it will always be true that $|\ell^{+}(R_i)| = \ell(R_i)$.

The bias, $B(\ell,W)$ can now be compared to zero. The closer it is to zero, the better the estimating procedure, if an unbiased estimating procedure is important.

The second consideration mentioned above is to obtain some measure of variability around the average loss. The usual procedure would be to make some sort of a variance calculation. For example,

$$\sum_{i=n_o+1}^{N} W_i[\ell(R_i) - A(\ell,W)]^2 \tag{19}$$

where $A(\ell,W)$ is the Average Loss, Eq. 14

The desirable property would be for 19 to be small. For our purposes, however, this is not very appropriate. As can be seen in Fig. 3 a small measure of variance would imply little chance of small losses as well as large losses. While the latter is to be avoided, the former is clearly desirable.

A measure of skewness* would hence be more appropriate than a measure of variance. Negative values of skewness, close to minus one, would imply that most residuals had small losses, hence small errors. A positive value would imply the opposite and would therefore detract from

*As defined by Cramér, Ref. 8, page 184, as μ_3/σ^3 where μ_3 is the third central moment and σ is the standard deviation.

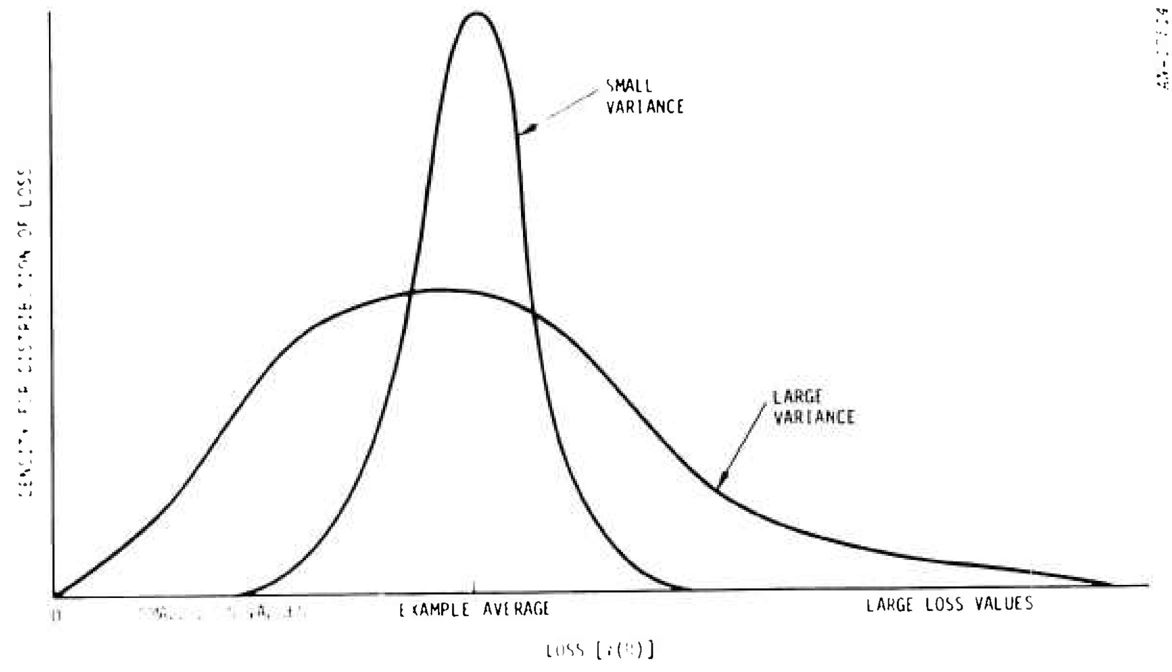

Figure 3(U). Variance Calculations and Average Loss an otherwisc good value of average loss. The skewness calculation for thjs application is given by

$$S_k(\ell,W) = \frac{\sum_{i=n_o+1}^{N} W_i[\ell(R_i) - A(\ell,W)]^3}{\left\{\sum_{i=n_o+1}^{N} W_i[\ell(R_i) - A(\ell,W)]^2\right\}^{3/2}} \tag{20}$$

where $A(\ell,W)$ is the Average Loss, Eq. 14

W_i are the weights used in Average Loss

$\ell(R_i)$ is the loss function used in Average Loss

The example given at the start of this section hypothesized two estimating procedures. Procedure A had $A(\ell,W) = 0.2$, while Procedure B had $A(\ell,W) = 0.25$. Answering the question of which one is preferable can be aided by calculating the measures just defined. Suppose that for Procedure A, $B(\ell,W) = 0.05$ and $S_k(\ell,W) = 0$. Then, if the equivalent measures for Procedure B were $B(\ell,W) = 0.1$ and $S_k(\ell,W) = 0.5$, the case for selecting A over B would be strengtłened. If, however, the measures for B were $B(\ell,W) = 0.01$ and $S_k(\ell,W) = -0.5$, the case for choosing A would be weaker. Procedure B is less biased and shows a large negative skewness which implies that losses smaller than the average were far more plentiful (or had more weight) in the sample (and hence we would hope more likely in the fu.ure) than were losses larger than the average. Procedure A, on the other hand, had zero skewness implying that large and small losses are equally likely.

4. Example Calculations

To familiarize the reader with the summarizations suggested in Sec. IV B 2 and the additional statistics defined in the last subsection (IV B 3), example calculations are made and presented for the computer program test data.

From the example used in the computer test run summarized in Tables 3 and 7 we have the following data (in the notation of Table 10).

TABLE 11

EXAMPLE DATA

$N = 13;\ n_o = 5;\ k = 3$

Sample Point		6	7	8	9	10	11	12	13
Residual	R_i	-11.7	-66.9	31.4	-9.9	4.7	-18.5	23.7	50.1
Subsample Size	S_i	5	6	7	8	9	10	11	12
Actual Cost	y_i	67	243	54	112	106	183	156	177

If the proportional error loss function is selected, then the average loss will be called average proportional error and is given by

$$\text{Average Proportional Error} = \sum_{i=6}^{13} W_i \frac{|R_i|}{y_i} \qquad \text{(after Eq. 10)}$$

The proportional error for each residual is given in Table 12.

TABLE 12

PROPORTIONAL ERROR

Sample Point	6	7	8	9	10	11	12	13
R_i	-11.7	-66.9	31.4	-9.9	4.7	-18.5	23.7	50.1
y_i	67	243	54	112	106	183	156	177
Proportional Error								
$\frac{\lvert R_i \rvert}{y_i}$	0.174	0.275	0.582	0.088	0.044	0.101	0.152	0.283

Three weighting schemes have been suggested in Table 10. These can be used to modifiy Eq. 10 as follows:

Average Proportional Error =
(Weight proportional to sample size)

$$\frac{1}{\sum\limits_{i=6}^{13}(S_i)} \sum_{i=6}^{13} S_i \frac{|R_i|}{y_i} \qquad (21)$$

Average Proportional Error =
(Weight proportional to degrees of freedom)

$$\frac{1}{\sum\limits_{i=6}^{13}(S_i - k)} \sum_{i=6}^{13} (S_i - k) \frac{|R_i|}{y_i} \qquad (22)$$

Average Proportional Error =
(Equal weight)

$$\frac{1}{8} \sum_{i=6}^{13} \frac{|R_i|}{y_i} \qquad (23)$$

All that remains is to substitute the values of S_i and k from Table 11 and $|R_i|/y_i$ from Table 12 and carry out the arithmetic. The results are given below. In this particular example, the weights do not greatly affect the average. All average proportional errors are around 20 percent.

Weight	Average Proportional Error
Proportional to Sample Size	0.202
Proportional to Degrees of Freedom	0.197
Equal Weight	0.212

The tendency in this example for larger proportional errors with predictions from smaller sample sizes can be seen by the fact that the equal weight measure gives the highest average proportional error while the degrees of freedom weighting scheme yields the lowest. These weighting schemes give the most and least weight to residuals calculated from small sample size predictions respectively.

Calculations for bias and skewness are made for the weighting scheme that is proportional to sample size only, i.e.,

$$W_i = \frac{S_i}{\sum_{i=6}^{13} S_i}$$

This should suffice to indicate how these measures are calculated.

The calculation for bias, Eq. 18, is very similar to those for average proportional error. All that should be done to Eq. 21, to obtain the signed loss (proportional error), is to remove the absolute value sign from R_i . Alternatively, one can use the proportional error from Table 12 and assign the sign of R_i from the same table. Thus, for data point 6, we have signed proportional error equals -0.174. The values to be averaged are given in Table 13 and the modified Eq. 18 for bias is given as Eq. 24. The value obtained for bias is 0.078. Note however, that the numbers of over- and underestimates are the same. The large error in estimating procurement 8 dominates the bias calculation.

$$\text{Bias}(\ell,W) = \frac{1}{\sum_{i=6}^{13} S_i} \sum_{i=6}^{13} S_i \frac{R_i}{y_i} \tag{24}$$

where ℓ is proportional error

W is the weighting scheme proportional to sample size

TABLE 13

BIAS CALCULATION VALUES

Sample Point		6	7	8	9	10	11	12	13
	S_i	5	6	7	8	9	10	11	12
Signed Proportional Error	$\frac{R_i}{y_i}$	-0.174	-0.275	0.582	-0.088	0.044	-0.101	0.152	0.283

By far the hardest measure to calculate is skewness. The modified version of the skewness equation, Eq. 20, is given below:

$$S_k(\ell,W) = \frac{\sum_{i=6}^{13} S_i\left(\frac{|R_i|}{|y_i|} - 0.202\right)^3}{\left\{\sum_{i=6}^{13} S_i\left(\frac{|R_i|}{|y_i|} - 0.202\right)^2\right\}^{3/2}} \qquad (25)$$

where $\frac{|R_i|}{y_i}$ is the proportional error

S_i is the sample size

0.202 is the average proportional error for the example

ℓ is proportional error

W is the weighting scheme proportional to sample size

The necessary data for the calculation can be obtained from Tables 11 and 12. A measure of skewness equal to 0.166 is obtained. Hence, the distribution shows some positive skewness, the large overestimate of sample point 8 outweighing the fact that 5 of the 8 proportional errors are less than the average.

It is hoped that the example calculations carried out in this section serve as a guide to help the reader make calculations of average loss, bias, and skewness for the loss function and weighting scheme best suited for his problem. It will be useful now to consider the possible directions of future work that might improve these measures.

5. Future Work

As pointed out at the beginning of this section, the arguments presented for the various residual summarizations and other measures have been heuristic in nature. This was due to the lack of an assumed underlying statistical model. The arguments are hence analogous to those that are used for various curve fitting schemes when a statistical model has not been assumed.

Several possible courses of action might be taken to either make the arguments for these statistics more rigorous or to derive better measures. Formal methods of nonparametric statistics might be useful in making more rigorous the comparison between estimating procedures A and B at the end of Sec. IV B 3. Another possibility is to explore the use of average loss for a ranking technique for several classes of candidate estimating procedures and their implied statistical models. This could be accomplished with the aid of Monte Carlo techniques. The probability of selecting the wrong estimating procedure, i.e., making an incorrect ranking, could be estimated.

The effort required to investigate these possibilities is certainly not trivial. In the meantime, the statistics suggested appear to be reasonable and should help the analyst to make choices between any candidate estimating procedures. In addition, several of the statistics identified, i.e., Average Proportional Error, Eq. 10, and Average Absolute Error, Eq. 17, have interpretations that are easy to communicate and can be used to give one a feeling of the estimating procedure's validity. They summarize the error which would have been present (on the average) if the candidate estimating procedure had been used to <u>predict</u> the cost in the past. Thus, these measures are good candidates for the desired measures identified in Sec. II C 1.

C. STATISTICS WHICH DEPEND ON A PARTICULAR ESTIMATING PROCEDURE

A final set of statistics can be calculated from the Historical Simulation output by making use of any statistical model assumptions that are usually associated with the particular cost estimating procedure under examination. An example is the multiple linear regression model, which is usually assumed when the cost estimating procedure of interest comprises a linear PER and a least squares technique[*] for estimating the parameters. Another example is the use of a multiplicative error term δ with a log-normal distribution[**] when the PER is given by

$$Y = aX_1^{b_1} X_2^{b_2} \cdot \cdot \cdot X_p^{b_p}$$

where Y stands for cost, $X_1, X_2, \ldots, X_p$ are the independent variables, and the technique is a least squares curve fit performed on

$$\log Y = \log a + b_1 \log X_1 + \cdot \cdot \cdot + b_p \log X_p$$

[*] In fact, the choice of the least squares technique can be viewed as a consequence of the multiple linear regression model assumption (for a linear PER) as the estimators obtained have some optimal properties. These properties are stated in the Gauss-Markov theorem. According to Ref. 7, page 387, this theorem states that "the least squares estimate in the class of unbiased, linear estimates, has a minimum variance property: the variances of its components are (simultaneously) smallest." In addition, they are maximum likelihood estimates.

[**] For additional information concerning this distribution see Ref. 7, page 89.

The statistical model in this case is

$$Y_i = aX_{1i}^{b_1} \cdot \cdot \cdot X_{pi}^{b_p} \epsilon_i$$

where $\log \epsilon_i$ is distributed normally with zero mean, variance equal to σ^2, and zero covariances.

From an operational point of view, statistics based on an assumed distribution are less versatile than the general data summarizations discussed in Sec. IV B. They are valid only if the assumed statistical model is valid.

However, these statistics are still worth examining. Since they are valid for any estimating procedures which utilize the same statistical assumptions, for example, the class of linear PERs (with least squares curve fit), they can be used to compare candidate estimating procedures in the class. However, these comparisons can also be made with the usual evaluation procedures, i.e., the usual regression statistics, and, hence, benefits gained using Historical Simulation do not include a comparison that cannot otherwise be directly made (Sec. II C, Property 2).

Another use for these statistics is to ascertain whether the statistical model and/or estimating procedure is valid. Does the Historical Simulation output fit in with the output that should be theoretically expected, assuming that the statistical model and estimating procedure assumptions are valid? If the output does not fit, then some of these assumptions must have been violated and hence the model should not be accepted.

Finally, statistics that are usually calculated (on the entire sample for the estimating procedure, can be derived for each of the Historical Simulation subsamples. For example, R^2 and standard error of the estimate can be calculated for each subsample, if the estimating procedure assumes a linear PER and a least squares curve fitting technique. These statistics can be used, in the traditional manner, to evaluate how well the estimating procedure is performing on the particular subsample. Would the model have been acceptable for the particular subsample? Would it have been rejected for a larger subsample?

Work accomplished to date on the development of these statistics has been confined to linear PERs, least squares estimating techniques, and the usual multiple linear regression model. This class of estimating procedures have been labeled Linear PER-Least Squares Procedures. The development of these statistics for this class of estimating procedures will have the added benefit that their study will more clearly define the relationship between the Historical Simulation output and the usual multiple linear regression evaluations.

To date the theoretical distribution of the Historical Simulation output--the predictions and residuals--have been determined. A goodness-of-fit test and a test to determine if there is bias present have been defined for a subset of the Historical Simulation output. Finally, several statistics have been identified that are useful in describing subsample fits. Each of these topics will be discussed in subsequent paragraphs, but first it will be useful (for clarity's sake) to define the Historical Simulation procedure (for multiple linear regression models) in matrix notation.

We are given a sample which consists of N P+1-tuples $(y_i , x_{i1} , x_{i2} , \ldots , x_{ip})$ for $i = 1, 2 , \ldots , N$. These P+1-tuples have been ordered in time.

The usual multiple linear regression hypothesis is given by

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix}$$

is a N × 1 column vector of observed y values

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{N1} & x_{N2} & & x_{Np} \end{pmatrix}$$

is a N × p+1 matrix of independent variable values (and a 1 for the constant multiplier)

$$\vec{\beta} = \begin{pmatrix} \beta_o \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}$$

is the p+1 × 1 column vector of model coefficients

and

$$\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_N \end{pmatrix}$$

is a N × 1 column vector of error terms

The matrix X and the vector $\vec{\beta}$ are assumed to be nonrandom; $\vec{\epsilon}$, on the other hand, is a normally distributed random vector with zero means, a constant variance σ^2, and zero covariances. That is

$$\left.\begin{aligned} E\epsilon_i &= 0 \\ \text{Variance}(\epsilon_i) &= \sigma^2 \\ \text{Covariance}(\epsilon_i,\epsilon_j) &= 0 \quad ; \quad i \neq j \end{aligned}\right\} i = 1, 2, \ldots, N$$

Let n_o be the minimum sample size that is greater than or equal to the smallest sample size necessary to carry out a linear regression analysis. Hence, $n_o \geq p + 2$. For any n, $n_o \leq n < N$, define the following partition of the X matrix by

$$X = \begin{bmatrix} X_1^{(n)} \\ \hline X_2^{(n)} \end{bmatrix} \begin{array}{l} n \text{ rows} \\ N\text{-}n \text{ rows} \end{array}$$

Also partition $\vec{Y}$ in a similar manner obtaining

$$\vec{Y} = \begin{pmatrix} \vec{Y}_1^{(n)} \\ \hline \vec{Y}_2^{(n)} \end{pmatrix} \begin{array}{l} n \text{ entries} \\ N\text{-}n \text{ entries} \end{array}$$

If time batches are ignored, the Historical Simulation Procedure can be defined as follows:

For each n, $n_o \leq n < N$

1. Make a least squares fit using $\vec{Y}_1^{(n)}$ and $X_1^{(n)}$ as the data base.

2. Obtain an estimating vector of $\vec{\beta}$. Denote this vector $\hat{\vec{\beta}}^{(n)}$

3. Use the resulting fit to make predictions of the remaining N-n data points. This can be denoted by

$$\hat{\vec{Y}}^{(n)} = \begin{pmatrix} \hat{y}_{n+1}^{(n)} \\ \hat{y}_{n+2}^{(n)} \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_{N}^{(n)} \end{pmatrix} = X_2^{(n)} \hat{\vec{\beta}}^{(n)}$$

where $\hat{y}_{n+k}^{(n)}$ is the prediction of y_{n+k} arrived at using a sample of size n.

4. Calculate the residuals by

$$D^{(n)} = \begin{pmatrix} d_{n+1}^{(n)} \\ d_{n+2}^{(n)} \\ \cdot \\ \cdot \\ \cdot \\ d_{n+k}^{(n)} \end{pmatrix} = \begin{pmatrix} \hat{y}_{n+1}^{(n)} - y_{n+1} \\ \hat{y}_{n+2}^{(n)} - y_{n+2} \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_{N}^{(n)} - y_{N} \end{pmatrix} = \hat{\vec{Y}}^{(n)} - \vec{Y}_2^{(n)}$$

where $d_{n+k}^{(n)}$ denotes the difference (residual) between the predicted $\hat{y}_{n+k}^{(n)}$ and the observed y_{n+k}.

1. Distribution of the Historical Simulation Predictions and Residuals

The form and distribution of this output has been summarized in Table 14, with derivations given in Appendix II. Several interesting results which can be observed from this table are discussed below.

TABLE 14

FORM AND DISTRIBUTION OF PREDICTIONS AND RESIDUALS

(Assuming the usual multiple linear regression model assumptions)

Notation	Prediction $\hat{y}^{(n)}_{n+k}$	Residual $d^{(n)}_{n+k}$
Sample point for which the prediction (residual) pertains	$n+k \;\; ; \;\; 0 < k \le N-n$	$n+k \;\; ; \;\; 0 < k \le N-n$
Subsample size used	n	n
Calculation	$\sum_{j=1}^{n} C_j^{n+k} y_j$	$\hat{y}^{(n)}_{n+k} - y_{n+k}$
Distribution	Normal	Normal
Expected value	$\dot{x}'_{n+k}\dot{\beta}$	0
Variance	$\sigma^2 C^{n+k}_{n+k}$	$\sigma^2\left(1 + C^{n+k}_{n+k}\right)$
Covariances (with)	$\hat{y}^{(m)}_{m+j}$	$d^{(m)}_{m+j}$
$m \le n$		
$m+j \le n$	$\sigma^2 C^{n+k}_{m+j}$	0^*
$m+j > n$		
and $m+j \ne n+k$		$\sigma^2 C^{n+k}_{m+j}$
and $m+j = n+k$		$\sigma^2\left(1 + C^{n+k}_{m+j}\right)$
$m > n$	(same as above with C^{m+j}_{n+k} replacing C^{n+k}_{m+j})	

where $\dot{x}'_i$ is a row vector equal to the ith row of the matrix X

$\dot{x}_i$ is a column vector equal to the ith row of the matrix X

y_j is the jth component of the vector $\dot{Y}$

$C_j^{n+k} = \dot{x}'_{n+k} S^{(n)^{-1}} \dot{x}_j$

$S^{(n)} = X_1'^{(n)} X_1^{(n)}$ where $X_1^{(n)}$ is the first n rows of X

and $X_1'^{(n)}$ is its transpose

$\dot{\beta}$ are the unknown parameters

and σ^2 is the variance of the error terms

The distribution of the prediction and residuals is normal. The predictions are unbiased in the sense that $E\hat{y}^{(n)}_{n+k} = \vec{x}'_{n+k}\beta = Ey_{n+k}$. Hence, the expected value of the residuals is zero. The variance of the residuals is related to the variance of the predictions in the sense that

$$\mathrm{VAR}(d^{(n)}_{n+k}) = \sigma^2 + \mathrm{VAR}(\hat{y}^{(n)}_{n+k})$$

This is a consequence of the fact that y_{n+k} is not used in the calculation of $\hat{y}^{(n)}_{n+k}$ and hence is independent of $\hat{y}^{(n)}_{n+k}$.

The residual covariances are related to the prediction covariances; in fact they are equal unless one of the points being predicted is not predicted from both subsamples, or the point being predicted is the same for both subsamples. In the first of these exceptions, i.e., when comparing $d^{(n)}_{n+k}$, $d^{(m)}_{m+j}$; $m+j \leq n$ or $n+k \leq m$, the covariance is zero. In the second exception, i.e., when comparing $d^{(n)}_{n+k}$, $d^{(m)}_{m+j}$; $m+j = n+k$, the calculation is similar to a variance calculation. The residual covariance is obtained by adding σ^2 to the prediction covariance. To summarize, then, we have for $m \leq n$

$$\mathrm{COV}\left(d^{(n)}_{n+k}, d^{(m)}_{m+j}\right) = \begin{cases} 0 & ; \quad m+j \leq n \\ \sigma^2 + \mathrm{COV}\left(\hat{y}^{(n)}_{n+k}, \hat{y}^{(m)}_{m+j}\right) & ; \quad m+j = n+k \\ \mathrm{COV}\left(\hat{y}^{(n)}_{n+k}, \hat{y}^{(m)}_{m+j}\right) & ; \quad \begin{matrix} m+j > n \text{ and} \\ m+j \neq n+k \end{matrix} \end{cases}$$

Another observation that can be made about the covariances is that they depend on the two subsample sizes m and n only through which S matrix to use. If $m \leq n$, then $S^{(n)}$ is used. (This is the only difference between the coefficients C^{n+k}_{m+j} and C^{m+j}_{n+k} in Table 14). The rule to follow is *always use the S matrix corresponding to the larger subsample size.*

If it is noted that $S^{(n)}$ is a function of $X^{(n)}$ (the data base used for the subsample size n fit) then the above result is not surprising. If $m \leq n$ then $X^{(n)}$ contains all the information that $X^{(m)}$ contains plus the extra rows. Hence $S^{(n)}$ will contain all the information available in $S^{(m)}$ plus more. Hence in deriving the covariance of two predictions or residuals, the information available in the larger subsample fit is required and includes the information available in the smaller subsample.

A final observation is that although the predictions and residuals are generally correlated (among themselves), there are some residuals which have zero covariance. In particular, if $m+j \leq n$, then $\mathrm{COV}\left(d_{n+k}^{(n)}, d_{m+j}^{(m)}\right) = 0$. In words this implies that the residual calculation for a particular sample point has zero covariance with any residual calclation based on a subsample which includes the specified sample point. The importance of this result lies in the fact that zero covariance implies independence when the random variables are normally distributed.* Hence $d_{n+k}^{(n)}$ and $d_{m+j}^{(m)}$ are independent if $m+j \leq n$. In particular then, the one-step residuals, i.e.,

$$d_{n_o+1}^{(n_o)}, d_{n_o+2}^{(n_o+1)}, \ldots, d_{n+1}^{(n)}, \ldots, d_N^{(N-1)}$$

are mutually independent. This fact will enable several statistical tests to be applied to the one-step residuals.

Before discussing these tests it is notationally convenient to redefine the one-step residuals as follows:

Let

$$r_n = \frac{d_{n+1}^{(n)}}{\sqrt{1 + C_{n+1}^{n+1}}}$$

*Zero covariance does not usually imply independence. The fact that the added condition of normalcy implies independence is discussed in Appendix II.

where $\quad C_{n+1}^{n+1} = \dot{x}'_{n+1} S^{(n)^{-1}} \dot{x}_{n+1}$ as defined in Table 14

Since the variance of $d_{n+1}^{(n)} = \sigma^2\left(1 + C_{n+1}^{n+1}\right)$ this transformation has the effect of giving the residuals $\vec{r} = (r_{n_o}, r_{n_o+1}, \ldots, r_{N-1})$ a common variance σ^2. Hence, we have that the residuals $\vec{r}$ are independent and normally distributed, with zero mean, and common variance σ^2. Alternatively $\vec{r}$ constitutes a random sample from a normal population with zero mean, and variance equal to σ^2.

2. Tests on the One-Step Residuals[*]

Two types of tests have been constructed for the one-step residuals. The first is a goodness-of-fit test which asks the question: *Do the one-step residuals appear to have been derived from a multiple linear regression model with the assumed linear PER?* The second addresses the question of bias and asks the question: *Do the one-step residuals appear to have zero mean (as they theoretically should)?*

The question of whether the model assumptions are satisfied has not been one of the central questions for theoretical statisticians. To be sure, a great body of knowledge has been built up around the closely related subject of hypothesis testing, but these tests are concerned with choosing between two states of nature, the null hypothesis and a specified[**] alternative hypothesis. The question we are asking can be placed in the hypothesis testing context. The null hypothesis H_0 is that $\vec{r}$ is a random sample from a normal population with zero (0) mean, and variance σ^2. Notationally this is given by:

$$H_0 : \vec{r} \overset{d}{=} N(0, \sigma^2 I)$$

[*] It should be pointed out that zero covariances were the requirements for these tests. Hence, the tests would appear to be equally applicable to all residuals if these residuals were orthogonalized. Pursuit of this topic is beyond the scope of the present work, however.

[**] The alternative hypothesis can be a class of alternative hypotheses such as, "The random vector $\vec{r}$ is from a normal distribution."

where I is the identity matrix of order $N-n_o$.

The alternative hypothesis is not specific, however. It is that $\vec{r}$ is not $N(0,\sigma^2 1)$. Hence the usual techniques for hypothesis testing, e.g., maximum likelihood, are not applicable. Fortunately, a few tests, called goodness-of-fit tests, have been devised to handle this question, but they are unfortunately not very powerful[*] against specific alternatives. Hence, if the question is to choose between two specified alternatives, a test built around these alternatives should be developed.

The two main types of goodness-of-fit-tests[**] are the Chi-Square test and tests that compare distribution functions. The Chi-Square test requires a partition of the sample and a comparison of the frequency of observations to the theoretical frequency. This test in general requires a large sample size and is therefore not very useable for the cost application.[***]

[*] There are two types of errors that can be made in a hypothesis testing problem. A type I error is made when H_0 is rejected and it was true. A Type II error is made when H_1 is rejected and it was true. Denote the probability for these two types of error by P_{H_0} (Reject H_0) and P_{H_1} (Reject H_1), respectively. The statement that goodness-of-fit tests are not very powerful against specific alternatives implies that in general there exists a hypothesis test for the specific alternative such that for a given P_{H_0} (Reject H_0), P_{H_1} (Reject H_1), using this other test, is less than P_{H_1} (Reject H_1) using the goodness-of-fit test. For a further discussion of this concept see Ref. 7, Chapter 7.

[**] See Ref. 7 Section 9.1 for a complete discussion.

[***] See Ref. 9, page 46.

Of the tests that compare distribution functions, the Kolmogorov-Smirnov (K-S) test is perhaps the most widely known. Its advantage over the Chi-Square test is that it appears to be more powerful and it is applicable for small sample sizes.* The test is also relatively simple. Tailored for our present application, it is outlined below.

Order the residual vector $\dot{r}$ from smallest to largest to obtain $r^{(1)}, r^{(2)}, \ldots, r^{(N-n_o)}$ where $r^{(j)}$ is the jth order statistic. Calculate the sample distribution function $F_{N-n_o}(r)$ by letting

$$F_{N-n_o}(r) = \frac{j}{N-n_o} \quad \text{for} \quad r^{(j)} \leq r < r^{(j+1)} \quad ; \quad j = 0, 1, \ldots, N-n_o \qquad (26)$$

where

$$r^{(0)} = -\infty \quad \text{and} \quad r^{(N-n_o+1)} = \infty$$

This sample distribution function is then compared to the theoretical distribution function under H_0, i.e., $F(r) = N(0,\sigma^2)$. The test statistic is defined by

$$D_{N-n_o} = \sup_{\text{all } r} \left| F_{N-n_o}(r) - F(r) \right|$$

that is, D_{N-n_o} is the largest absolute difference between the two distribution functions. It can be shown that the distribution of D_{N-n_o} is not dependent on the distribution of $F(r)$.** Values of the distribution of D_{N-n_o} are tabulated in most statistics books (see Ref. 7, Table VI) and rejection values are given based on the significance level of the test desired (i.e., the probability of a Type I error allowable). Thus all that remains is to determine D_{N-n_o}.

* See Ref. 9, page 51.

** Ref. 7, page 300.

For our application, this constitutes nothing more than comparing $F_{N-n_o}(r)$ to $F(r)$ at the end points of the steps in $F_{N-n_o}(r)$. Thus D_{N-n_o} will be the maximum of the numbers

$$\left|\frac{j-1}{N-n_o} - F(r^{(j)})\right| \text{ and } \left|\frac{j}{N-n_o} - F(r^{(j)})\right| \quad ; \quad j = 1, \ldots, N-n_o \qquad (27)$$

The quantities $F(r^{(j)})$ are easy to determine for a given σ^2. By using any table of the normal distribution, one merely looks up the value of the percentile of the Normal distribution function for $r^{(j)}/\sigma$. A problem arises however by what value to use for σ^2.

Three candidates are presented and discussed in Appendix III and the square of the Standard Error of the Estimate which is obtained in the usual regression analysis--from the fit on the entire sample N--is selected. The choice was based on the fact that it was the most efficient estimator and that unlike the other candidates, it does not depend directly on the residuals in $\vec{r}$. Furthermore, it is the estimate of the variance that is normally used in a regression analysis. The estimator is denoted $\hat{\sigma}^2$ and the equation for calculating it is given by:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - (P+1)} \qquad (28)$$

where y_i is the actual cost of the ith procurement

$\hat{y}_i$ is the estimated cost of the ith procurement (obtained from a regression analysis of the entire sample)

and P is the number of independent variables in the PER

The K-S test is then valid as long as we define the null hypothesis, H_0, as "$\vec{r}$ is a random sample from a $N(0,\sigma^2)$ distribution." The test is not ncessarily valid for the wider null hypothesis of H_0 defined as

"$\vec{r}$ is a random sample from $N(0,\sigma^2)$ with σ^2 estimated by $\hat{\sigma}^2$." There are indications, however, that if rejection takes place then rejection would also take place if σ^2 were known (Ref. 9, page 60). In addition, Darling, in Ref. 10, has described some conditions under which variations of the K-S test are valid for the wider hypotheses. How the present application fits into his work has yet to be determined. Further research will have to be done on extending the current application to this wider hypothesis.

The second test proposed in this section addresses the question of bias in the vector of residuals $\vec{r}$. In particular it is a hypothesis test given by

H_0: $\vec{r}$ is a random sample from a $N(0,\sigma^2)$ population

H_1: $\vec{r}$ is a random sample from a $N(\mu,\sigma^2)$ population

where $\mu \neq 0$.

The test statistic is derived* by the use of a likelihood ratio test. The test statistic has a t-distribution with $N-n_0-1$ degrees of freedom and is given by

$$t = \frac{(N-n_0-1)^{1/2}\bar{r}}{S_r} \qquad (29)$$

where $\bar{r}$ is the sample mean of $\vec{r}$, i.e.,

$$\bar{r} = \frac{\sum_{i=n_0}^{N-1} r_i}{N-n_0} \qquad (30)$$

*For derivation see Ref. 7, page 320.

and S_r^2 is the sample variance given by

$$S_r^2 = \frac{\sum_{i=n_o}^{N-1} (r_i - \bar{r})^2}{N-n_o} \qquad (31)$$

The test is conducted as follows:

1. Determine the significance level α, i.e., what probability is the analyst willing to withstand of rejecting H_0 when it is true?
2. From a t-table,* obtain the value of the $\alpha/2$ and $1 - \alpha/2$ percentiles of the t-distribution with $N-n_o - 1$ degrees of freedom. Label these $t_{\alpha/2}$ and $t_{1-\alpha/2}$. Note that only one value need be obtained, as $t_{\alpha/2} = -t_{1-\alpha/2}$.
3. Calculate t from Eq. 29.
4. If $t_{\alpha/2} \leq t \leq t_{1-\alpha/2}$, then H_1 is rejected and no apparent bias is present (at the α-significance level).
5. If $t < t_{\alpha/2}$ or if $t > t_{1-\alpha/2}$, then H_0 is rejected and there is significant bias present (at the α significance level).

Another way of stating this test is to ask the question: is $\bar{r}$ significantly different from zero? If so, H_0 should be rejected and bias is present.

An example of the use of these tests is given below. Again, the data base used will be the one that was used in the computer test run. Values of $\vec{r}$ are obtained from output block 7, Table 20, Appendix I. These are given in Table 15.

*Available in any statistics book, such as Ref. 7.

TABLE 15

ONE-STEP ADJUSTED RESIDUALS FROM TEST RUN*

Sample Point	Sample Predicted From	Adjusted Residual
6	5	-10.63
7	6	-21.71
8	7	25.95
9	8	-8.038
10	9	2.889
11	10	-15.67
12	11	21.02
13	12	37.721

*Source, STAT 1, Output Block 7, Table 20, Appendix I.

To apply the K-S test, we must first order the adjusted residuals from smallest to largest. Then the residuals are divided by $\hat{\sigma}$, i.e., the standard error of the estimate from sample size 13. From the computer test run, last output block 5 (Table 20, Appendix 1) $\hat{\sigma} = 21.6$. By using tables of the Standard Normal Distribution, these latter quantities are converted to percentiles of the Standard Normal distribution (equivalent to obtaining their cummulative distribution function value). These operations are summarized in Table 16, columns 2-4.

These percentiles are to be compared to the endpoint values of the steps in the sample distribution function given in Eq. 27. Since there are eight sample points, the values of the sample distribution function will jump by one-eighth. The appropriate endpoint values are given in columns 5 and 6 of Table 15.

The maximum differences between the percentiles (in column 4) and the endpoints (in columns 5 and 6) are calculated for each sample point.

TABLE 16

K-S TEST CALCULATIONS

Sample Point	Ordered Adjusted Residuals	Divided by Standard Error of Estimate	Percentile or Normal Population	Compare to End Points of Sample Distribution Function		Maximum Difference For Point
7	-21.71	-1.005	0.157	0	0.125	0.157
11	-15.67	-0.725	0.234	0.125	.250	·.125
6	-10.63	-0.492	.311	.250	.375	·.125
9	-8.038	-0.372	.355	.375	.500	.145
10	2.889	0.134	.553	.500	.625	·.125
12	21.02	0.973	.835	.625	.750	.210
8	25.95	1.201	.885	.750	.875	.135
13	37.721	1.746	.960	.375	1.000	·.125

These are shown in column 7. The K-S statistic for sample size (number of residuals) 8, D_8 , is then the maximum value in column 7. In the example under discussion $D_8 = 0.21$. This value is within acceptable limits, as D_8 would have to be greater than 0.358 for rejection at as high a significance level (probability of making a Type I error) as 0.2. Hence the regression model and linear PER cannot be rejected.

The calculation for bias is performed by first calculating the sample mean, Eq. 30, and sample variance, Eq. 31, for the residuals $\vec{r}$, (column 3, Table 15). These calculations resulted in values of $\bar{r} = 3.94$ for the sample mean and $S_r = 20.7$ for the standard deviation.

The t-statistic is then given by Eq. 29 as

$$t = \frac{(N-n_o-1)^{1/2}\bar{r}}{S_r} \tag{29}$$

Hence, in our case

$$t = \frac{(7)^{1/2}3.94}{20.7} = 0.504$$

This is not a significant t-value (7 degrees of freedom) for any reasonable significance level. As an example, if $\alpha = 0.2$, i.e., the 0.2 significance level, then the rejection limits would be ±1.42. The value of t obtained above is not even close to being outside of this range. Hence the Historical Simulation results do not indicate bias in the model.

Even though the values of these two statistics are insignificant for the test run data, there will be times when they are significant, and yet the usual regression statistics would seem reasonable. To illustrate this point consider the theoretical example portrayed in Fig. 4.

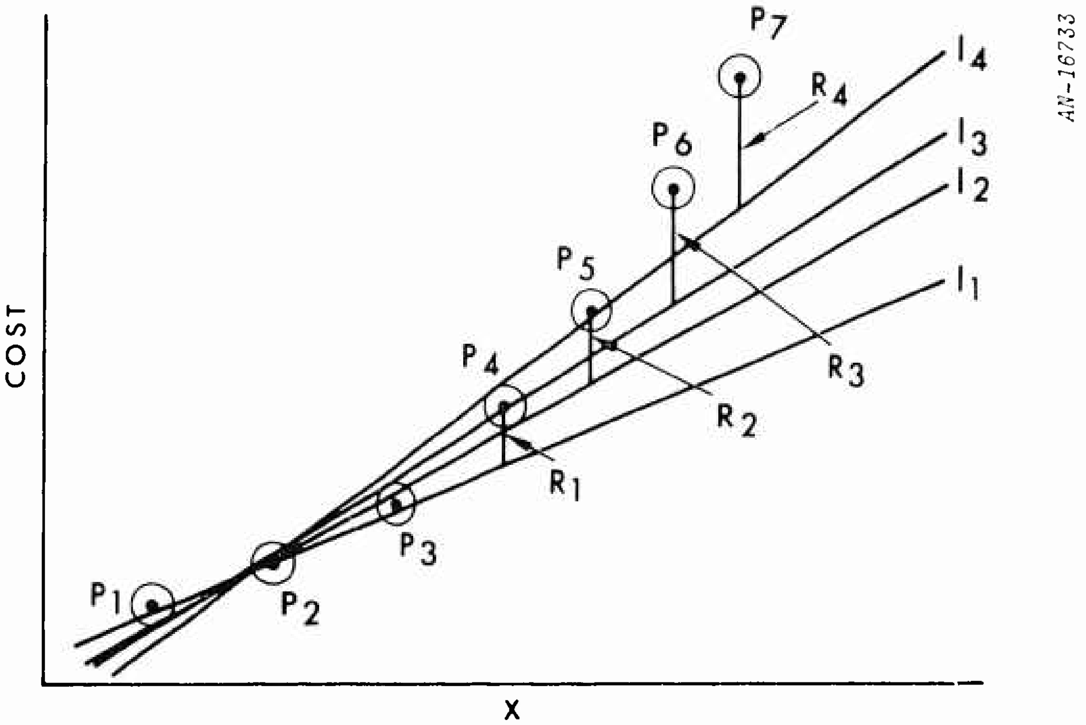

Figure 4(U). Theoretical Example

In the example, the candidate estimating procedure is

$$\text{Cost} = a + bX$$

and a least squares curve-fitting technique is used to pick the parameters. The time sequencing of the data is the same as an ordering on the values of X , i.e., larger values came later.

At the first stage of Historical Simulation the first three data points (P_1, P_2 , and P_3) are used to fit a line ℓ_1 . The estimate of P_4 would be low by the amount R_1 . At the next stage of Historical Simulation, line ℓ_2 would be derived using as the data base points P_1, P_2, P_3 and P_4 . The estimate of P_5 derived from ℓ_2 would be low by R_2 . The process is continued deriving lines ℓ_3 from data points P_1 through P_5 , and ℓ_4 from P_1 through P_6 . The estimates of P_6 (from ℓ_3) and P_7 (from ℓ_4) are low by R_3 and R_4 , respectively. Thus all the predictions obtained were low. This implies that the K-S statistic and t statistic will most likely be significant,[*] and hence the estimating procedure would not be accepted by Historical Simulation.

Looking at each of the lines, however, it does not seem that the fit (to the data they were derived from) is too bad In fact, ℓ_4 would probably be accepted as a good model for the first six data points, using statistics based on regression theory.[**] Hence the model would be accepted using the regression theory statistics while it would be rejected using Historical Simulation.

Of course, in this case, a simple plot of the data would convince an analyst that he has the wrong model (it should be exponential rather than linear). This, however, is a consequence of a two-dimensional problem (Cost and X) in which plots can be made and our illustration could be drawn. The analyst will not have the luxury of such plots when working with more than one independent variable, and an extension of this example to a multiple independent variable model can readily be made (without a figure, however).

[*] While the significance of the t-statistic depends on the magnitude of the residuals and how close together they are, the fact that all residuals are negative will usually lead to rejection of the zero mean hypothesis. In regard to the K-S test, all negative residuals implies a K-S statistic value greater than 0.5. This is significant for four residuals at the 0.2 significance level and if the process in the example continues for seven residuals. The results will be significant at the 0.05 level

[**] It should be noted that there is another technique, called Time Sequence Plot of the residuals (Ref. 6, page 68), which for the example being discussed would result in a sequencing of residuals from the usual regression analysis that would indicate a lack of fit. However, the consequences of retaining the model (in this example), i.e., the likelihood of underestimates, are more apparent when processed by Historical Simulation. Furthermore, even though residual plots should be analyzed whenever a least squares curve fit is made, the fact is that such examinations of residuals are often forgotten.

It should be noted that the two tests discussed in this section are very different. The test for bias assumes that the underlying model is normal and all that is being tested is if the mean is zero. The Kolmogorov-Smirnov test, on the other hand, asks whether or not the distribution is normal, with mean 0 and variance $\hat{\sigma}^2$. Both tests address the question of model validity, however, as the residuals should theoretically come from an $N(0,\sigma^2)$ population.

It is expected that other tests can be constructed for the one-step residuals. In addition to the above and extensions of them to tests applied to all the residuals (after some orthogonalization), it will be worthwhile to develop two hypothesis tests where H_1 is some other candidate estimating procedure. If this alternative is also a linear PER, with the assumed multiple linear regression model, then such tests should be relatively easy to construct. If the alternative is a nonlinear PER, then the appropriate statistical distribution will have to be identified and the distribution of the Historical Simulation predictions and residuals will have to be derived. Then the question of devising tests can be addressed. Needless to say, this last group of tests will take considerable effort.

3. Comparison Statistics

The last set of statistics that have been identified* are some of the usual regression statistics for each of the subsample fits in Historical Simulation. (They have nothing to do with the prediction and residual output of Historical Simulation.) These can be used in the usual manner to see how well the estimating procedure is doing on each of the subsamples. They also can be directly compared to like statistics on the entire sample.

*Example values for the test run can be seen in output blocks 5, Table 20, Appendix I.

The first set of measures are best summarized as Measures of Fit. Two such measures are calculated on each subsample for which parameters are estimated. These measures are given below:

$$\left.\begin{array}{l}\text{Standard Error}\\ \text{of the Estimate}\end{array}\right\} = \text{SEE} = \left[\frac{1}{m-k}\sum_{i=1}^{m}(A_i - \tilde{P}_i)^2\right]^{1/2} \qquad (32)$$

$$\left.\begin{array}{l}R^2 \text{ or Coefficient}\\ \text{of Determination}\end{array}\right\} = \frac{\sum_{i=1}^{m}(\tilde{P}_i - \bar{A})^2}{\sum_{i=1}^{m}(A_i - \bar{A})^2} \qquad (33)$$

where m = subsample size

A_i = actual cost of the ith object

$\tilde{P}_i$ = estimate of the ith object (Fit)

k = number of parameters to be estimated in the PER

and $\bar{A}$ is the average of the A_i's, i.e.,

$$\bar{A} = \frac{1}{m}\sum_{i=1}^{m} A_i$$

These measures are not at all related to the predictions calculated from the CER that is derived by fitting the curve to the subsample. They merely describe how good the fit was. In theory, if the process satisfies the statistical assumptions, SEE^2 should be converging to the true variance σ^2, and R^2 should be converging to

$$1 - \frac{\sigma^2}{\frac{1}{m-1}\sum_{i=1}^{m}(A_i - \bar{A})^2}$$

Therefore, as the sample size increases through the Historical Simulation evaluation, we should see this convergence (although for most cost applications the number of samples fitted will probably be too small). In practice it is desirable for σ^2 to be small, and therefore a good fit is represented by a SEE close to zero and an R^2 close to one.

Another set of fit statistics are the t-statistics for each estimated coefficient of the linear model. These are the statistics that are usually used to see if a coefficient is significantly different from zero.

Given that these coefficients are all different from zero in a usual regression run (i.e., for the entire sample), it may turn out that they are not significant for all of the subsample fits processed in Historical Simulation. It seems reasonable that once the subsample size was large enough for all to be significant, then they should remain significant. If not, one might begin to question the value of retaining the independent variable that corresponds to the occasionally significant coefficient.

Note also that the fact that a particular coefficient is not significant for early data bases brings into question the relevance of that data base to the current prediction problem. It may be useful to try estimating procedures which ignore this early data. Such a procedure would be one that estimates the parameters using, say, only the last 6 data points in time.

In summary, these statistics are useful in seeing how the fit is improving as the sample size grows. They do not, however, pertain to the main output of Historical Simulation, i.e., the predictions and residuals. They should shed some light on any anomalies present in this latter output, however, and may be useful in suggesting new candidate estimating procedures.

This concludes the discussion of the uses of the output from Historical Simulation and the work to date on its development. Next i. seems appropriate to summarize the advantages and current limitations of Historical Simulation and indicate the direction of possible future research. These topics are discussed in the next section.

V. CONCLUSION AND RECOMMENDATIONS FOR FUTURE EFFORT

In concluding this report, it will be useful to summarize the limitations and advantages of Historical Simulation as it is currently envisioned. This section will be itself concluded with some recommendations for future work which would hopefully shed further light on some of the limitations noted and expand on the work already completed.

A. CURRENT LIMITATIONS

That limitations exist is not always bad, as the following discussion will show. However, there are areas where the development of Historical Simulation is far from complete and the attendant limitations are a real problem. These limitations, as the author currently sees them, are discussed below.

1. Lack of a Single Way to Interpret the Output

Whether this is really a limitation or not is open to question. It would certainly be more convenient if one summarization could answer all our questions about a cost estimating procedure's reliability and validity. But this type of convenience is not even present in the use of regression theory, as can be seen from the several statistics that must be calculated (e.g., R^2, standard error of estimate, and prediction intervals). Furthermore, the lack of such a convenient data summarization has the effect of forcing the analyst to examine the residuals (Table 9), something that should be done anyway.

2. Lack of Ability to Uniquely Specify the Minimum Sample Size, n_o, for Historical Simulation

As discussed in Sec. III C, the specification of a minimum sample size is not a trivial problem. To be sure, there is a lower bound (depending on the number of PER parameters) below which the value of n_o cannot be defined, but this lower bound is just a starting point in the specification of n_o.

If too low a value is specified, there may not be enough degrees of freedom for initial predictions to be very meaningful. On the other hand, too large a value of n_o greatly diminishes the Historical Simulation output. Each additional sample point included in the initial subsample deletes a row from the prediction and residual output matrices. Hence the analyst must specify n_o to be the smallest number for which the estimating procedure, if valid, will have enough information from which to make reasonable predictions.

3. Loss of Information in the Data Summarizations and Statistics Derived in Sec. IV

Due to a lack of independence, summarizations suggested in Secs. IV B and IV C have only made use of one residual calculation for each sample point, usually the one-step residuals $d_{n+1}^{(n)}$. Hence a great deal of information goes unused. Further research should be initiated to try to incorporate the unused information in the recommended summarizations and tests. Some nonparametric statistical techniques might prove useful for the summarizations that do not depend on a particular estimating procedure while orthogonalization techniques could be applied to the residuals that are based on the Linear PER-Least Squares procedures.

4. Lengthy Output Time Requirements for the Time Share Computer Model

The output time requirements for operation of the computer model on the GE time sharing service seem undesirably long. Thirty-one minutes of terminal time was required for the test run, Table 20, Appendix I. There are no inherent reasons for this. It is probably possible to write the program or program output format more efficiently. Another possibility is to convert the program to a non-time-sharing machine with more efficient output. Since the program has been written in FORTRAN, this latter course should pose few problems.

5. Lack of Application in the Development of Actual CERs

The usefulness of Historical Simulation will ultimately be decided by the analyst. Several ways of using the output have been suggested in Sec. IV. Their value in selecting between several candidate cost estimating procedures and in hypothesizing new cost estimating procedure candidates can only be evaluated through their attempted application. From this process, it is expected that new uses of the output will be created and perhaps some of the suggested uses discarded.

Some examples of the application of Historical Simulation are given in Volume 2 of this report. They, however, do not serve to remove this limitation, as a much greater exposure is required to fully understand the practical worth of Historical Simulation. Furthermore the author lacks the necessary understanding of either the data base or the example aircraft programs to fully utilize the Historical Simulation output.

6. Lack of a Precise Understanding as to the Situations for Which Historical Simulation Will be More Valuable Than Regression Techniques

Insights into the relationship between these two techniques have been achieved in Sec. IV C and Appendixes II and III. The fact that there are situations in which Historical Simulation will be more valuable is clear (see Fig. 4, pg. 71). Also it seems clear that Historical Simulation provides a greater visibility (e.g., the identification of questionable sample points or the demonstration of successful extrapolations) than the usual regression techniques, even when the conclusions reached by the two techniques are the same.

However, a precise understanding of all the possible situations for which one of the techniques is more valuable will probably never be reached. This fact leads to the final limitation.

7. Historical Simulation is Not the Ultimate Answer, Merely Another Tool

Used in conjunction with such traditional methods as regression theory, Historical Simulation should improve the quality of our CERs. Furthermore, no technique, including Historical Simulation, will ever remove the necessity for the analyst. He is, in fact, an integral part of the evaluation procedure. He must choose candidate estimating procedures, examine the output tables, choose loss functions and weighting schemes, etc. Hence the best that can be done is to provide him with as many useful tools as possible to best perform his analysis.

B. ADVANTAGES

Several unique advantages of the Historical Simulation procedure have been identified throughout this report. These are summarized below.

1. Historical Simulation Can Compare a Wider Class of Cost Estimating Procedures Than the Usual Regression Techniques

Section III demonstrated the ability of Historical Simulation to evaluate any cost estimating procedure.

2. Historical Simulation Provides an Easy-to-Communicate Summary Statistic Useful for Describing the Accuracy of a Prediction

This summary statistic is average proportional (or absolute) error or one of its weighted forms. While it does not summarize all of the Historical Simulation output it does describe how well the cost estimating procedure would have predicted if it had been used in the past to make predictions of the now historical data.

3. Historical Simulation Provides a View Independent of the Usual Regression Theory Approach

This independent view is a consequence of the fact that Historical Simulation evaluates the ability of a candidate cost estimating procedure to predict the future from the past. Historical Simulation does not depend on how well the candidate cost estimating procedure fits the data. Consequences of this are

1. An independent view of CERs derived from stepwise regression
2. Additional information to help hypothesize a new cost estimating procedure candidate
3. Exposure of questionable sample points which do not fit in with the prior data base in terms of information content for parameter estimation and in terms of simulated predictions.
4. A demonstration of the candidate estimating procedure's ability to extrapolate from historical data to make predictions.
5. The possibility of uncovering errors in an estimating procedure's formulation which would not be uncovered by the usual regression statistics, e.g., Fig. 4, page 71.

C. RECOMMENDATIONS FOR FUTURE EFFORT

This report has described the work accomplished to date on the development of Historical Simulation. It is the author's opinion that the procedure has been developed sufficiently and offers enough advantages for it to be usefully applied by those analysts in industry and government involved in the development of CERs.

However, as we have pointed out in this section, there are limitations that should be examined so that the Historical Simulation procedure can be more fully developed and hence more meaningfully applied. The future effort required should proceed along two distinct paths, one theoretical, the other applied.

On the theoretical side, three classes of problems can be identified for future investigation.

1. Incorporation of more of the residual output into the suggested statistics and tests: Examples were discussed in limitation 3.

2. Determination of the probability of selecting the wrong estimating procedure when using the Sec. IV B summary statistics (e.g., average loss) for ranking: Monte Carlo techniques applied to the usually assumed statistical models for the candidate estimating procedures might be used for this investigation.

3. Determination of the theoretical distribution of the Historical Simulation residuals for estimating procedures other than Linear PER-Least Sr'ares Procedures: Exponential PERs, Eq. 2, utilizing a log-linear curve fitting technique are examples of alternative estimating procedures that should be explored.

On the applied side, the use of Historical Simulation in the development of CERs should be encouraged. This work should be carried out by individual analysts engaged in the development of CERs, for only they will have the knowledge of their data base and of the pnysical makeup of the class of procurements under investigation necessary to interpret the Historical Simulation output and to hypothesize new cost estimating procedure candidates. Of course, reporting of the successes, failures, or extensions of the Historical Simulation procedure which are discovered in specific applications should also be encouraged.

APPENDIX I

COMPUTER PROGRAM DESCRIPTION--LINEAR PER-LEAST SQUARES CLASS EXAMPLE

A. GENERAL REMARKS

In this appendix the relationship of the Historical Simulation procedure to an estimating procedure is described in detail by examining a computer program developed for Historical Simulation. This program has been written in FORTRAN for the G.E. Time Sharing Service, MARK I; the main program is listed in Table 17. In describing this program the flow diagram of Fig. 5 will be followed. The figure is divided into two parts. On the left, under the title of Main Program, are those calculations which are not dependent upon a particular estimating procedure. To these operations the calculations peculiar to a given estimating procedure are added, as portrayed in the right side of Fig. 5, under the heading Estimation Procedure.

In theory, a set of operations should be supplied for each estimating procedure being tested, but fortunately it appears that these operations can be more generally written around classes of estimating procedures. As an example, a set of operations written for estimating procedures which use the least squares fit technique and a (multivariate) linear PER will be discussed. The multiple linear regression model is usually assumed for this class of estimating procedures and the class has been referred to as the Linear PER-Least Square Procedures.

In writing the program the operations under Estimation Procedure were organized into four subroutines. Since this organization will probably be useful for any estimating procedure (that can be automated), it will be useful to document it here. The subroutine names, appropriate box numbers from Fig. 5, and table numbers for a complete listing of the programs are listed in the following table.

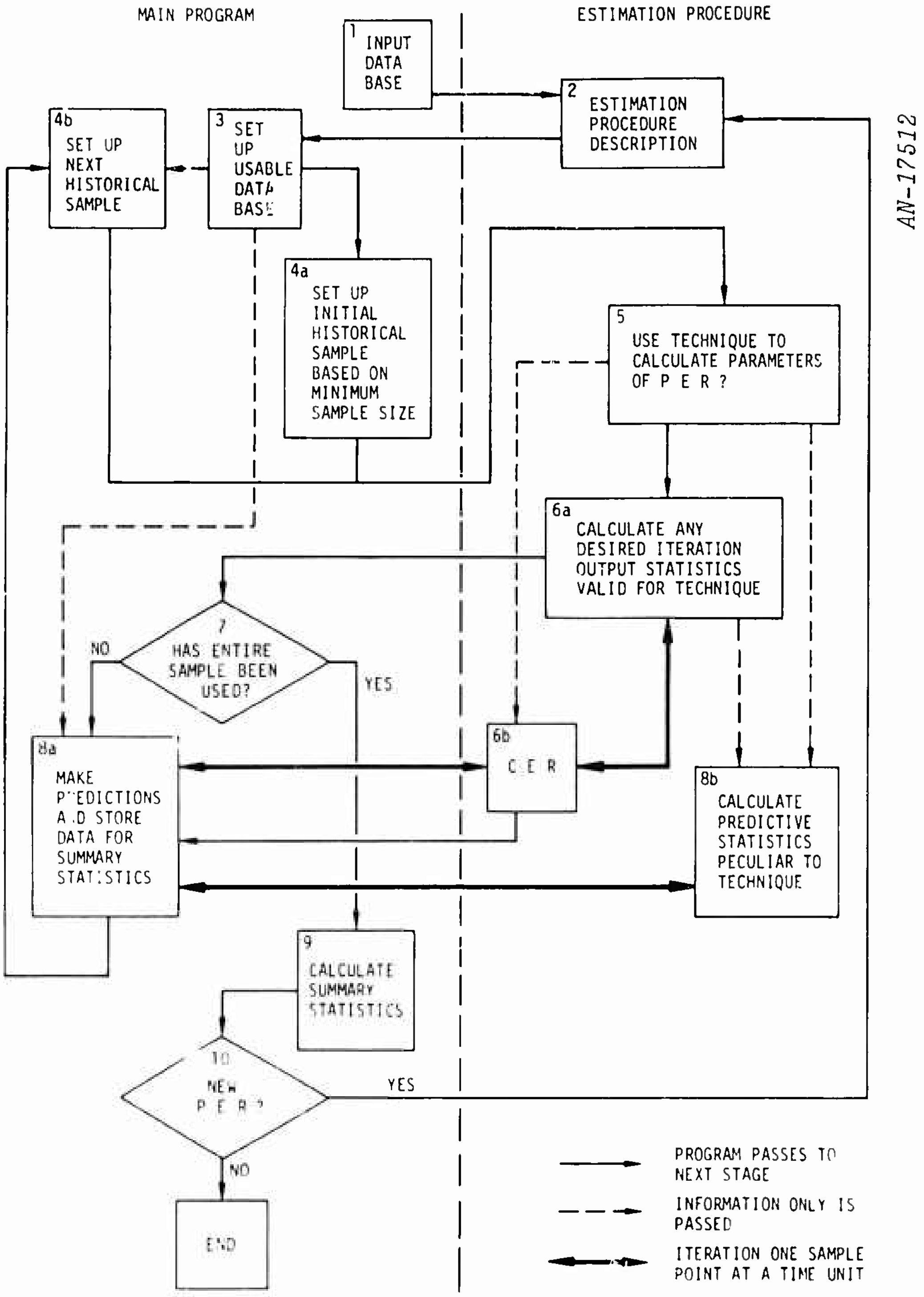
MAIN PROGRAM
ESTIMATION PROCEDURE
1 INPUT DATA BASE
2 ESTIMATION PROCEDURE DESCRIPTION
3 SET UP USABLE DATA BASE
4b SET UP NEXT HISTORICAL SAMPLE
4a SET UP INITIAL HISTORICAL SAMPLE BASED ON MINIMUM SAMPLE SIZE
5 USE TECHNIQUE TO CALCULATE PARAMETERS OF P E R ?
6a CALCULATE ANY DESIRED ITERATION OUTPUT STATISTICS VALID FOR TECHNIQUE
7 HAS ENTIRE SAMPLE BEEN USED?
NO
YES
6b C E R
8a MAKE PREDICTIONS AND STORE DATA FOR SUMMARY STATISTICS
8b CALCULATE PREDICTIVE STATISTICS PECULIAR TO TECHNIQUE
9 CALCULATE SUMMARY STATISTICS
10 NEW P E R ?
YES
NO
END
PROGRAM PASSES TO NEXT STAGE
INFORMATION ONLY IS PASSED
ITERATION ONE SAMPLE POINT AT A TIME UNIT
AN-17512

Figure 5(U). Relationship of Historical Simulation and the Estimation Procedure

Subroutine Name	Operation Number (from Figure 5)	Program Listing, Table Number
DESCP	2, 10	18
TECH	5, 6a	19
EST	6b	18
SST	8b	18

In Table 20, an example is given of the Historical Simulation output (using a Linear PER-Least Squares procedure). The data do not represent any real data base and the reader is therefore cautioned about drawing conclusions. Examples with aircraft and helicopter data are given in Vol. 2 of this report.

As the program is being discussed, reference will frequently be made to Tables 17 through 20. The contents of Tables 17 through 19 will be referred to by line number. The contents of Table 20 will be reterenced by output group (numbers 1-7 in the left margin).

TABLE 17

MAIN PROGRAM

```
HISS


Ø     SFILE RUNDAT,TOTDAT
1     COMMON IDV(6),NIV,EMU(7),RDATA(42,7),VAL(6),
2     +AINV(7,7),AVG(7),SSE
3     DIMENSION DATA(42,7),REST(42,7),ITBAT(42),NORDE(42),NW(42)
4     EQUIVALENCE (DATA,REST)
9C    SAMPLE INPUT AND TIME BATCHES
12    READ(1),NUMV,NUMS,OUT,NUMT
13    NUMV1=NUMV + 1
14    READ(1),(ITBAT(I),I=1,NUMT)
15    15  READ (1),(NW(I),I=1,NUMS)
17    17  READ(2),((DATA(I,J),J=1,NUMV1),I=1,NUMS)
18    READ (2),NORD
19    IF(NORD - 1) 29
2Ø    READ(2),(NORDE(I),I=1,NUMS)
21    REWIND 2
22    READ(2),((RDATA(I,J),J=1,NUMV1),I=1,NUMS)
25    DO 25, I=1,NUMS
26    DO 25, J=1,NUMV1
27    DATA(NORDE(I),J) = RDATA(I,J)
28    25  RDATA(I,J)=Ø
29    29  REWIND 2
3Ø    IF (3-OUT)4Ø
31    PRINT,
33    PRINT,"                         SAMPLE DATA"
37    PRINT 28,
41    28 FORMAT(4HSMP.,9X,9HACT. VAL.,25X,19HVAL. OF INDEP. VAR.,
42    +/3HNO.,27X,8HX1,(4,7),9X,8HX2,(5,8),9X,8HX3,(6,9))
53    DO 32,I=1,NUMS
57    32 PRINT 33, I,DATA(I,NUMV+1),(DATA(I,J),J=1,NUMV)
61    33 FORMAT(I3,1X,4F17.2,F36.2,2F17.2,F36.2,2F17.2)
62    OUT=OUT+3
65C   OBTAIN TECHNIQUE DESCRIPTION
69    4Ø  CALL DESCP(MSAMS,NUMV)
7Ø    NIV1=NIV+1
73C   REDUCE DATA FOR THIS TECHNIQUE
85    DO 6Ø,I=1,NUMS
86    RDATA(I,NIV1)=DATA(I,NUMV1)
87    DO 6Ø,J=1,NIV
88    K=IDV(J)
89    6Ø RDATA(I,J) = DATA(I,K)
1Ø1C  SET UP FIRST SAMPLE
1Ø5   IF(MSAMS-NUMS)85,85
1Ø9   PRINT,"                    SAMPLE TOO SMALL"
113   STOP
117   85 NUMS1 = Ø
121   DO 9Ø, I=1,NUMT
125   ORCSS-NUMS1+ITBAT(I)
129   IF (ORGSS-MSAMS)9Ø,1ØØ,1ØØ
133   9Ø NUMS1=NUMS1+ITBAT(I)
137C  SET UP SAMPLES
```

TABLE 17 (cont'd)

MAIN PROGRAM

```
HISS    CONTINUED


141    100 DO 400,JA=1,NUMT
143    PRINT; PRINT; PRINT,
145    NUMS1 = NUMS1+ITBAT(JA)
149    IF(NUMT-JA)110,120,130
153    110 PRINT,"          SAMPLE SET UP WRONG"
157    STOP
161    120 PRINT,"           ENTIRE SAMPLE USED"
177    130 PRINT 136,NUMS1
181    136  FORMAT(14HSAMPLE SIZE = ,I3)
188    X=1.
189    CALL TECH(OUT,NUMS1)
193    PRINT,
197C   TEST TO SEE IF DONE
201    IF(NUMT-JA) 110,410
209C   SET UP PREDICTION OUTPUT; CALC. STAT.
216    NUMS11=NUMS1+1
217    DO 200,I=NUMS11,NUMS
220    REST(I,5)=RDATA(I,NIV1)
221    DO 167,J=1,NIV
225    167 VAL(J) = RDATA(I,J)
226    REST(I,1)=REST(I,5);REST(I,2)=EST(X);REST(I,6)=REST(I,2)
233    CALL SST(REST(I,5),REST(I,6),NUMS1)
239    REST(I,3)=REST(I,2)-REST(I,1)
240    200  REST(I,4)=REST(I,3)/REST(I,1)
250    IF(5-OUT) 400
251    PRINT,
252    PRINT,"                     PREDICTIONS"
253    PRINT,
254    PRINT 185,
255    DO 400,I=NUMS11,NUMS
256    PRINT  190,(REST(I,J),J=1,6)
257    400  CONTINUE
260    185  FORMAT(6X,6HACTUAL,5X,8HESTIMATE,3X,10HDIFFERENCE,2X,
261    +9HPROP.ERR.,4X,6HSTAT.1,6X,6HSTAT.2)
265    190  FORMAT(6F12.3)
277C   FINAL OUTPUT
281    410 PRINT,
285    PRINT,"                    FINAL OUTPUT"
289    NSW = 0; APE = 0.; BIA = 0.; SK1 = 0.; SK2 = 0.
293    PRINT 185,
296    ORGSS1 = ORGSS+1
297    DO 430 I=ORGSS1,NUMS
301    PRINT 190,(REST(I,J),J=1,6)
303    APE = APE +ABSF(REST(I,4))*NW(I)
304    435  BIA = BIA + REST(I,4))*NW(I)
309    430  NSW = NSW + NW(I)
310    PRINT;PRINT;
318    APE = APE/NSW;  BIA = BIA/NSW
319    DO 437 I = ORGSS1,NUMS
320    RED = ABSF(REST(I,4)) - APE
322    SK1 = SK1 + NW(I)*RED**2
323    437  SK2 = SK2 + NW(I)*RED**3
325    PRINT,"AVE. PROPORTIONAL ERROR =",APE
326    PRINT,"BIAS                    =",BIA
327    PRINT,"SKEWNESS                =",SK2/SK1**1.5
350    GO TO 17
351    END
900    SUSE HISST
936    SUSE HISS3
938    SUSE HISS1
940    SOPT SIZE
```

TABLE 18

HISS 3

THREE SUBROUTINES

```
HISS3


335  SUBROUTINE DESCP(MSAMS,NUMV)
339  COMMON IDV(6),NIV,EMU(7),RDATA(42,7),VAL(6),
340  +  AINV(7,7),AVG(7),SSE
347C SPECIFY MINIMUM SAMPLE SIZE AND INDEP. VARIABLE
351 READ(1), NIV
359 MSAMS=NIV+2
363 DO 518 J=1,NIV
366 READ (1),IDV(J)
367 518 IF(NUMV-IDV(J)) 545
375C OUTPUT TECHNIQUE
379 PRINT; PRINT,
387  PRINT,"                      LINEAR PER - LEAST SQUARES"
391 PRINT 535,NIV,(IDV(J),J=1,NIV)
395 535 FORMAT(2ØHNO. OF INDEP. VAR.= ,I1,3X,18HVARIABLE
399 + NOS. ARE ,9I3)
4Ø3 RETURN
407 545 PRINT,"         UNDEFINED VARIABLE CALL IN DESCP"
411 STOP
415 END
42Ø FUNCTION EST(X)
424  COMMON  IDV(6),NIV,EMU(7),RDATA(42,7),VAL(6),
425  + AINV(7,7),AVG(7),SSE
432 EST=EMU(NIV+1)
436 DO 555, J=1,NIV
44Ø 555 EST=EST+EMU(J)*VAL(J)
444 RETURN
448 END
58Ø SUBROUTINE  SST(S1,S2, NUMS1)
581  COMMON IDV(6),NIV,EMU(7),RDATA(42,7),VAL(6),
582  + AINV(7,7),AVG(7),SSE
584C TTEST OF NEW POINTS
585  SID= (1.+1./FLOATF(NUMS1))*SSE**2
586 DO 59Ø, I=1,NIV
587 DO 59Ø, J=1,NIV
59Ø 59Ø SID=SID+(VAL(I)-AVG(I))*AINV(I,J)*(VAL(J)-AVG(J))
591  S1 = SSE*(S2-S1)/SID**.5
595C  NO SECOND STATISTIC
596  S2 = Ø.
598 RETURN
599 END
```

TABLE 19

HISST

SUBROUTINE TECH

HISST

```
6ØØ SUBROUTINE TECH(OUT,NUMS1)
6Ø1  COMMON  IDV(6),NIV,EMU(7),RDATA(42,7),VAL(6),
6Ø2  + AINV(7,7),AVG(7),SSE
6Ø8  NIV1=NIV+1
61ØC  CALCULATE ARITHMETIC MEANS
615 DO 63Ø,I=1,NIV1
62Ø AVG(I)=Ø.
621 DO 625,J=1,NUMS1
625 625 AVG(I)=AVG(I)+RDATA(J,I)
63Ø 63Ø AVG(I)=AVG(I)/NUMS1
65ØC CLEAR CROSS PRODUCT MATRIX AND VECTOR
651  DO 655 I=1,NIV1
652 EMU(I)=Ø.
653  DO 655 J=1,NIV1
655 655 AINV(I,J)=Ø.
67ØC FORM CROSS PRODUCT MATRIX AND VECTOR
671 DO 68Ø I=1,NIV
673 DO 677 J=1,NUMS1
674 EMU(I)=EMU(I)+(RDATA(J,NIV1)-AVG(NIV1))*(RDATA(J,I)-AVG(I))
676 DO 677, K=I,NIV
677 677 AINV(I,K)=AINV(I,K)+(RDATA(J,I)-AVG(I))*(RDATA(J,K)-AVG(K))
678 DO 68Ø,K=I,NIV
68Ø  68Ø  AINV(K,I) = AINV (I,K)
7ØØC INVERT MATRIX
7Ø2  CALL MTINV(D,ID)
7Ø3 IF (ABSF(D)-.ØØØØ1)863
7Ø5C SET UP ESTIMATOR VECTOR
71Ø  EMU(NIV1)=AVG(NIV1)
712 DO 72Ø,I=1,NIV
72Ø 72Ø EMU(NIV1)=EMU(NIV1)-AVG(I)*EMU(I)
721  EVAR = Ø.; EBAR = Ø.; EEX = Ø.
722 IF(-XABSF(OUT-4)) 741
725C ESTIMATE VARIANCE; OUTPUT ESTIMATES
727 PRINT,"                                  SAMPLE"
728 PRINT 73Ø,
73Ø 73Ø FORMAT(5X,6HACTUAL,12X,8HESTIMATE,1ØX,1ØHDIFFERENCE,5X,
731 +13HPROPOR. ERROR)
74Ø  74Ø  FORMAT(E13.5,2E19.5,E17.5)
741  741  DO 755,I=1,NUMS1
743 DO 745,J=1,NIV
```

TABLE 19 (cont'd)

HISST

SUBROUTINE TECH

HISST   CONTINUED

```
745 745 VAL(J)=RDATA(1,J)
747 E=EST(D);A=RDATA(1,NIV1);B=E-A;C=B/A
748 IF(-XABSF(OUT-4))75Ø
749 PRINT 74Ø, A, E, B, C
75Ø  75Ø  EBAR = EBAR + (A - AVG(NIV1))**2
751  EVAR = EVAR +B**2
755 755EEX=EEX+(E-AVG(NIV1))**2
76ØC CALCULATE OUTPUT STATISTICS
78Ø 78Ø PRINT,
79Ø SSE=(EVAR/(NUMS1-NIV1))**.5
8ØØ  8ØØFORMAT(2ØHSTD. ERROR OF EST. =,E15.5,1ØX,
8Ø1 +12HR SQUARED = ,E15.5)
8Ø5C SET UP VAR - COV MATRIX
8Ø6 DO 81Ø, K=1,NIV
81Ø 81Ø VAL(K) = Ø.
82Ø DO 825, K=1,NIV1
824 DO 825,J=1, NIV
825 825 VAL(K) = VAL(K)-AINV(K,J)*AVG(J)
827 VAL(NIV1)=1/NUMS1
828 DO 83Ø, K=1,NIV
83Ø 83Ø VAL(NIV1)=VAL(NIV1)-VAL(K)*AVG(K)
835 DO 84Ø, K=1,NIV
836 DO 837, J=1,NIV
837 837 AINV(K,J) = AINV(K,J)*SSE**2
84Ø  84Ø  AINV(NIV1,K)=VAL(K)*SSE**2;AINV(K,NIV1)=AINV(NIV1,K)
845 AINV(NIV1,NIV1)=VAL(NIV1)*SSE**2
846  PRINT,
847  PRINT,"                         SUB SAMPLE STATISTICS"
848  PRINT,
849  PRINT  8ØØ, SSE,EEX/EBAR
85Ø  PRINT,
851 PRINT 86Ø,NUMS1-NIV1
852 DO 853,I=1,NIV
853 853 PRINT 861,I,EMU(I),EMU(I)/AINV(I,I)**.5
855  PRINT 862,EMU(NIV1)
857  RETURN
86Ø 86ØFORMAT(8HVARIABLE,11X,9HPARAMETER,15X,6HT TEST,14X,6HD.F. =,I3)
861 861 FORMAT(3X,I2,3X,3F2Ø.5)
862  862  FORMAT(8HCONSTANT,F2Ø.5)
863 863 PRINT,"DETERM=ZERO"
865  RETURN
87Ø END
```

TABLE 20

EXAMPLE OUTPUT

```
                              SAMPLE DATA
    SMP.          ACT. VAL.                          VAL. OF INDEP. VAR.
    NO.                           X1,(4,7)          X2,(5,8)          X3,(6,9)
     1             95.00          1996.00            178.00            153.00
     2             31.00           967.00            204.00            144.00
     3             60.00          2414.00            217.00            149.00
     4             82.00          4418.00            201.00            144.00
     5             25.00           852.00            172.00            107.00
1    6             67.00          2072.00            215.00            136.00
     7            243.00         10408.00            221.00            177.00
     8             54.00          2643.00            258.00            160.00
     9            112.00          3786.00            211.00            172.00
    10            106.00          3335.00            280.00            203.00
    11            183.00          6374.00            305.00            196.00
    12            156.00          7092.00            294.00            187.00
    13            177.00         10304.00            280.00            167.00

                  LINEAR PER-LEAST SQUARES
2   NO. OF INDEP. VAR. = 2    VARIABLE NOS. ARE    1    3

3   SAMPLE SIZE =    5
                                   SAMPLE
          ACTUAL              ESTIMATE          DIFFERENCE          PROPOR. ERROR
        .95000E+02           .67981E+02        -.27019E+02          -.28441E+00
        .31000E+02           .50156E+02         .19156E+02           .61793E+00
4       .60000E+02           .69160E+02         .91599E+01           .15267E+00
        .82000E+02           .86038E+02         .40380E+01           .49243E-01
        .25000E+02           .19666E+02        -.53344E+01          -.21338E+00

                              SUB SAMPLE STATISTICS

    STD. ERROR OF EST. =      .24755E+02          R SQUARED =          .67448E+00

5   VARIABLE            PARAMETER             T TEST                  D.F. = 2
       1                    .01040          1.08495
       2                    .79175          1.05986
    CONSTANT            -73.90994

                              PREDICTIONS

         ACTUAL        ESTIMATE       DIFFERENCE        PROP.ERR.       STAT.1         STAT.2
         67.000         55.311         -11.689           -.174         -10.6[illegible]      .000
        243.000        174.447         -68.553           -.282        [illegible]            .000
         54.000         80.250          26.250            .486        [illegible]            .000
6       112.000        101.636         -10.364           -.093        [illegible]            .000
        106.000        121.490          15.490            .146        [illegible]            .000
        183.000        147.546         -35.454           -.194        [illegible]            .000
        156.000        147.886          -8.114           -.052         -3.[illegible]        .000
        177.000        165.448         -11.552           -.065         -3.[illegible]        .000
```

TABLE 20 (cont.)

EXAMPLE OUTPUT

```
3 { SAMPLE SIZE =     6
                                        SAMPLE
       ACTUAL                ESTIMATE              DIFFERENCE           PROPOR. ERROR
     .95ØØØE+Ø2            .69535E+Ø2            -.25465E+Ø2           -.268Ø6E+ØØ
     .31ØØØE+Ø2            .51863E+Ø2             .2Ø863E+Ø2            .673ØØE+ØØ
4    .6ØØØØE+Ø2            .7Ø859E+Ø2             .1Ø859E+Ø2            .18Ø98E+ØØ
     .82ØØØE+Ø2            .88Ø48E+Ø2             .6Ø479E+Ø1            .73754E-Ø1
     .25ØØØE+Ø2            .22364E+Ø2            -.26364E+Ø1           -.1Ø546E+ØØ
     .67ØØØE+Ø2            .57332E+Ø2            -.96682E+Ø1           -.1443ØE+ØØ

                            SUB SAMPLE STATISTICS

   STD. ERROR OF EST. =        .21124E+Ø2             R SQUARED =           .64993E+ØØ

   VARIABLE              ARAMETER                  T TEST                 D.F. =   3
5     1                    .Ø1Ø49                1.28247
      2                    .7647Ø                1.2Ø391
   CONSTANT              -68.39237

                            PREDICTIONS

        ACTUAL        ESTIMATE     DIFFERENCE     PROP.ERR.      STAT.1        STAT.2
       243.ØØØ        176.Ø9Ø       -66.91Ø        -.275        -21.71Ø         .ØØØ
        54.ØØØ         81.672        27.672         .512         22.6Ø8         .ØØØ
       112.ØØØ        1Ø2.333        -9.167        -.Ø82         -6.474         .ØØØ
6      1Ø6.ØØØ        121.8Ø9        15.8Ø9         .149          7.621         .ØØØ
       183.ØØØ        148.321       -34.679        -.19Ø        -16.695         .ØØØ
       156.ØØØ        148.967        -7.Ø33        -.Ø45         -3.32Ø         .ØØØ
       177.ØØØ        167.352        -9.648        -.Ø55         -3.125         .ØØØ

3 { SAMPLE SIZE =     7
                                        SAMPLE
       ACTUAL                ESTIMATE              DIFFERENCE           PROPOR. ERROR
     .95ØØØE+Ø2            .6896ØE+Ø2            -.26Ø4ØE+Ø2           -.2741ØE+ØØ
     .31ØØØE+Ø2            .44254E+Ø2             .13254E+Ø2            .42754E+ØØ
     .6ØØØØE+Ø2            .73595E+Ø2             .13595E+Ø2            .22658E+ØØ
4    .82ØØØE+Ø2            .1Ø582E+Ø3             .23817E+Ø2            .29Ø45E+ØØ
     .25ØØØE+Ø2            .16Ø97E+Ø2            -.89Ø29E+Ø1           -.35612E+ØØ
     .67ØØØE+Ø2            .58322E+Ø2            -.8678[illegible]E+Ø1           -.12953E+ØØ
     .243ØØE+Ø3            .23596E+Ø3            -.7Ø[illegible]E+Ø1           -.28987E-Ø1

                            SUB SAMPLE STATISTICS

   STD. ERROR OF EST. =        .[illegible]E+Ø2             R SQUARED =           .9[illegible]36E+ØØ

   VARIABLE              PARAMETER                  T TEST                 D.F. =   4
5     1                    .Ø1[illegible]                [illegible]
      2                    [illegible]                1.1Ø[illegible]
   CONSTANT              -[illegible]

                            PREDICTIONS

        ACTUAL        ESTIMATE     DIFFERENCE     PROP.ERR.      STAT.1        STAT.2
        [illegible].ØØØ        8[illegible].441        31.441         [illegible]         25.95Ø         .ØØØ
       [illegible].ØØØ        114.298         2.298         .Ø21          1.749         .ØØØ
       [illegible].ØØØ        128.124        22.124         .2Ø9         1Ø.[illegible]         .ØØØ
6      [illegible].ØØØ        [illegible].398        -[illegible].6Ø2        -[illegible]         -3.528         .ØØØ
       [illegible].ØØØ        183.857        27.857         .179         2Ø.181         .ØØØ
       [illegible].ØØØ        [illegible]        5Ø.Ø[illegible]         .283         3[illegible].649         .ØØØ
```

TABLE 20 (cont.)

EXAMPLE OUTPUT

```
3 { SAMPLE SIZE =    8
                                             SAMPLE
        ACTUAL                  ESTIMATE               DIFFERENCE            PROPOR. ERROR
      .95000E+02               .60224E+02             -.34776E+02           -.36606E+00
      .31000E+02               .36792E+02              .57921E+01            .18684E+00
      .60000E+02               .67113E+02              .71127E+01            .11855E+00
4 {   .82000E+02               .10501E+03              .23008E+02            .28059E+00
      .25000E+02               .21808E+02             -.31915E+01           -.12766E+00
      .67000E+02               .55887E+02             -.11113E+02           -.16587E+00
      .24300E+03               .23475E+03             -.82505E+01           -.33953E-01
      .54000E+02               .75418E+02              .21418E+02            .39663E+00

                           SUB SAMPLE STATISTICS

  STD. ERROR OF EST. =           .22286E+02           R SQUARED =          .92572E+00

  VARIABLE               PARAMETER                  T TEST                D.F. =   5
5 {   1                    .01977                 5.16064
      2                    .34353                  .58173
  CONSTANT               -31.79035

                           PREDICTIONS

        ACTUAL     ESTIMATE    DIFFERENCE    PROP. ERR.      STAT.1        STAT.2
       112.000     102.134       -9.866        -.088         -8.038         .000
       106.000     103.869       -2.131        -.020         -1.165         .000
6 {    183.000     161.536      -21.464        -.117        -14.648         .000
       156.000     172.637       16.637         .107         12.691         .000
       177.000     229.258       52.258         .295         37.294         .000


3 { SAMPLE SIZE =    9
                                             SAMPLE
          ACTUAL                  ESTIMATE               DIFFERENCE            PROPOR. ERROR
        .95000E+02               .62293E+02             -.32707E+02           -.34429E+00
        .31000E+02               .38348E+02              .73480E+01            .23703E+00
        .60000E+02               .68574E+02              .85741E+01            .14290E+00
        .82000E+02               .10507E+03              .23069E+02            .28133E+00
4 {     .25000E+02               .19473E+02             -.55266E+01           -.22106E+00
        .67000E+02               .56112E+02             -.10888E+02           -.16251E+00
        .24300E+03               .23573E+03             -.72710E+01           -.29922E-01
        .54000E+02               .77952E+02              .23952E+02            .44355E+00
        .11200E+03               .10545E+03             -.65493E+01           -.58476E-01

                             SUB SAMPLE STATISTICS

  STD. ERROR OF EST. =            .20607E+02           R SQUARED =          .92555E+00

  VARIABLE               PARAMETER                   T TEST               D.F. =   6
5 {   1                    .01933                  5.27881
      2                    .45003                   .97170
  CONSTANT               -45.15223

                           PREDICTIONS

           ACTUAL     ESTIMATE    DIFFERENCE    PROP ERR.      STAT.1        STAT.2
          106.000     110.682        4.682         .044          2.889         .000
6 {       183.000     166.287      -16.713        -.091        -12.465         .000
          156.000     176.119       20.119         .129         16.260         .000
          177.000     229.218       52.218         .295         37.266         .000
```

TABLE 20 (cont.)

EXAMPLE OUTPUT

```
3 { SAMPLE SIZE =   10
                                              SAMPLE
           ACTUAL                 ESTIMATE               DIFFERENCE          PROPOR. ERROR
         .95000E+02              .61592E+02              -.33408E+02          -.35166E+00
         .31000E+02              .37856E+02               .68557E+01           .22115E+00
         .60000E+02              .68175E+02               .81745E+01           .13624E+00
         .82000E+02              .10540E+03               .23398E+02           .28534E+00
4        .25000E+02              .20817E+02              -.41830E+01          -.16732E+00
         .67000E+02              .56285E+02              -.10715E+02          -.15992E+00
         .24300E+03              .23582E+03              -.71777E+01          -.29538E-01
         .54000E+02              .77053E+02               .23053E+02           .42691E+00
         .11200E+03              .10422E+03              -.77804E+01          -.69468E-01
         .10600E+03              .10778E+03               .17826E+01           .16817E-01

                                    SUB SAMPLE STATISTICS

    STD. ERROR OF EST. =           .19110E+02              R SQUARED =           .92613E+00

    VARIABLE                  PARAMETER                   T TEST                D.F. =   7
5        1                        .01957                7.27017
         2                        .39968                1.40131
    CONSTANT                    -38.62356

                                    PREDICTIONS

              ACTUAL          ESTIMATE        DIFFERENCE      PROP.ERR.        STAT.1         STAT.2
             183.000          164.464          -18.536          -.101          -15.670          .000
6            156.000          174.919           18.919           .121           16.232          .000
             177.000          229.790           52.790           .298           38.056          .000
```

TABLE 20 (cont.)

EXAMPLE OUTPUT

```
3 { SAMPLE SIZE =   11
                                         SAMPLE
       ACTUAL                 ESTIMATE             DIFFERENCE          PROPOR. ERROR
     .95ØØØE+Ø2              .625Ø8E+Ø2           -.32492E+Ø2          -.342Ø2E+ØØ
     .31ØØØE+Ø2              .37831E+Ø2            .68314E+Ø1           .22Ø37E+ØØ
     .6ØØØØE+Ø2              .68869E+Ø2            .88691E+Ø1           .14782E+ØØ
     .82ØØØE+Ø2              .1Ø615E+Ø3            .24148E+Ø2           .29449E+ØØ
     .25ØØØE+Ø2              .17849E.Ø2           -.715Ø8E+Ø1          -.286Ø3E+ØØ
4 {  .67ØØØE+Ø2              .55878E+Ø2           -.11122E+Ø2          -.166ØØE+ØØ
     .243ØØE+Ø3              .24Ø52E+Ø3           -.24813E+Ø1          -.1Ø211E-Ø1
     .54ØØØE+Ø2              .78666E+Ø2            .24666E+Ø2           .45576E+ØØ
     .112ØØE+Ø3              .1Ø7Ø4E+Ø3           -.49643E+Ø1          -.44324E-Ø1
     .1Ø6ØØE+Ø3              .11294E+Ø3            .69421E+Ø1           .65492E-Ø1
     .183ØØE+Ø3              .16975E+Ø3           -.13247E+Ø2          -.72388E-Ø1

                              SUB SAMPLE STATISTICS

  { STD. ERROR OF EST. =          .18715E+Ø2              R SQUARED =          .93468E+ØØ

5 { VARIABLE               PARAMETER                 T TEST                  D.F. = 8
       1                      .Ø198Ø                7.54797
       2                      .47853                1.81979
    CONSTANT               -5Ø.22ØØ9

                            PREDICTIONS

       ACTUAL       ESTIMATE     DIFFERENCE     PROP.ERR.       STAT.1        STAT.2
6 {   156.ØØØ       179.66Ø        23.66Ø          .152         21.Ø2Ø          .ØØØ
      177.ØØØ       233.675        56.675          .32Ø         41.525          .ØØØ

3 { SAMPLE SIZE =  12
                                         SAMPLE
       ACTUAL                 ESTIMATE             DIFFERENCE          PROPOR. ERROR
     .95ØØØE+Ø2              .62Ø46E+Ø2           -.32954E+Ø2          -.34689E+ØØ
     .31ØØØE+Ø2              .38348E+Ø2            .73479E+Ø1           .237Ø3E+ØØ
     .6ØØØØE+Ø2              .68246E+Ø2            .82456E+Ø1           .13743E+ØØ
     .82ØØØE+Ø2              .1Ø432E+Ø3            .22315E+Ø2           .27213E+ØØ
     .25ØØØE+Ø2              .1959ØE+Ø2           -.54Ø95E+Ø1          -.21638E+ØØ
4 {  .67ØØØE+Ø2              .5589ØE+Ø2           -.1111ØE+Ø2          -.16582E+ØØ
     .243ØØE+Ø3              .23358E+Ø3           -.9415ØE+Ø1          -.38745E-Ø1
     .54ØØØE+Ø2              .77546E+Ø2            .23546E+Ø2           .436Ø4E+ØØ
     .112ØØE+Ø3              .1Ø477E+Ø3           -.72346E+Ø1          -.64595E-Ø1
     .1Ø6ØØE+Ø3              .11ØØ2E+Ø3            .4Ø182E+Ø1           .379Ø8E-Ø1
     .183ØØE+Ø3              .16498E+Ø3           -.16Ø23E+Ø2          -.98486E-Ø1
     .156ØØE+Ø3              .17467E+Ø3            .18674E+Ø2           .1197ØE+ØØ

                              SUB SAMPLE STATISTICS

  { STD. ERROR OF EST. =           .18985E+Ø2             R SQUARED =          .92976E+ØØ

5 { VARIABLE               PARAMETER                 T TEST                  D.F. =  9
       1                      .Ø1912                7.3847Ø
       2                      .44755                1.687Ø6
    CONSTANT               -44.58316

                              PREDICTIONS

        ACTUAL       ESTIMATE     DIFFERENCE     PROP.ERR.       STAT.1        STAT.2
6 {    177.ØØØ       227.122        5Ø.122          .283         37.721          .ØØØ
```

TABLE 20 (cont.)

EXAMPLE OUTPUT

```
                ENTIRE SAMPLE USED
3   SAMPLE SIZE =   13
                                         SAMPLE
          ACTUAL               ESTIMATE               DIFFERENCE        PROPOR. ERROR
        .95ØØØE+Ø2            .64234E+Ø2            -.3Ø766E+Ø2        -.32386E+ØØ
        .31ØØØE+Ø2            .42177E+Ø2             .11177E+Ø2         .36Ø55E+ØØ
        .6ØØØØE+Ø2            .68374E+Ø2             .83743E+Ø1         .13957E+ØØ
        .82ØØØE+Ø2            .9715ØE+Ø2             .1515ØE+Ø2         .18475E+ØØ
        .25ØØØE+Ø2            .17Ø56E+Ø2            -.79442E+Ø1        -.31777E+ØØ
        .67ØØØE+Ø2            .54744E+Ø2            -.12256E+Ø2        -.18293E+ØØ
4       .243ØØE+Ø3            .21334E+Ø3            -.29661E+Ø2        -.122Ø6E+ØØ
        .54ØØØE+Ø2            .78946E+Ø2             .24946E+Ø2         .46196E+ØØ
        .112ØØE+Ø3            .1Ø471E+Ø3            -.72932E+Ø1        -.65118E-Ø1
        .1Ø6ØØE+Ø3            .117Ø4E+Ø3             .11Ø35E+Ø2         .1Ø411E+ØØ
        .183ØØE+Ø3            .161Ø4E+Ø3            -.21961E+Ø2        -.12ØØ1E+ØØ
        .156ØØE+Ø3            .16681E+Ø3             .1Ø811E+Ø2         .693Ø4E-Ø1
        .177ØØE+Ø3            .2Ø539E+Ø3             .28388E+Ø2         .16Ø38E+ØØ


                           SUB SAMPLE STATISTICS

    STD. ERROR OF EST. =       .216Ø2E+Ø2           R SQUARED =        .9Ø936E+ØØ

    VARIABLE              PARAMETER              T TEST                 D.F. =  1Ø
5        1                  .Ø1593              6.889Ø4
         2                  .62944              2.22171
    CONSTANT              -63.86653


                         FINAL OUTPUT
              ACTUAL     ESTIMATE    DIFFERENCE   PROP.ERR.     STAT.1       STAT.2
              67.ØØØ      55.311      -11.689       -.174      -1Ø.631        .ØØØ
             243.ØØØ     176.Ø9Ø      -66.91Ø       -.275      -21.71Ø        .ØØØ
              54.ØØØ      85.441       31.441        .582       25.95Ø        .ØØØ
             112.ØØØ     1Ø2.134       -9.866       -.Ø88       -8.Ø38        .ØØØ
             1Ø6.ØØØ     11Ø.682        4.682        .Ø44        2.889        .ØØØ
             183.ØØØ     164.464      -18.536       -.1Ø1      -15.67Ø        .ØØØ
7            156.ØØØ     179.66Ø       23.66Ø        .152       21.Ø2Ø        .ØØØ
             177.ØØØ     227.122       5Ø.122        .283       37.721        .ØØØ


    AVE. PROPORTIONAL ERROR =          .2Ø27
    BIAS                    =          .Ø779
    SKEWNESS                =          .1662
```

B. PROGRAM INPUT

Data for the program are stored in two data files (for purposes of compilation economy) called RUNDAT and TOTDAT. TOTDAT consists of the data base, i.e., the physical and performance characteristics and cost of the historical procurements. Each row of the data matrix corresponds to one procurement (Table 21). Each of the procurements, NUMS in number, has associated with it a cost and a value for each physical and performance characteristic. If there are NUMV characteristics, then there will be NUMV+1 entries for each procurement, and hence there will be NUMV+1 times NUMS numbers in the procurement data base.

TABLE 21

DATA BASE ARRANGEMENT IN TOTDAT

	Procurement Number	Physical or Performance Characteristic Number						
		1	2	3	4	....	NUMV	Cost
Oldest	1	DATA ENTRIES						
	2							
	3							
	4							
	.							
	.							
	.							
Newest	NUMS							

NUMV: number of Physical and Performance Characteristics

NUMS: number of Procurements

For Historical Simulation, the procurements must be ordered in time, with the oldest in the first line of data. If this is the order of the data in TOTDAT, then enter a zero after the data for the last procurement. This value is used by an indicator variable NORD which leaves the data base alone when it equals zero.

If, however, the data base has a different order, let the value of NORD equal one. Follow this by NUMS numbers, one for each procurement indicating the transformation necessary to order the data in time.

The data used in the test run are given in Table 22. The data base is contained in the first 13 lines, 101-113, one for each procurement. There are four entries in each line, as there are values for three independent variables and the cost for each procurement. Hence, NUMS = 13 and NUMV = 3 foι the test run. The next entry, line 120, gives NORD a value of one, hence the data will be reordered. The new ordering is given in line 121. The first row of data, line 101, will become row 2, the second line will become row 4 and the fourth line will become row 1. All other rows will remain the same.

RUNDAT contains the remaining data arranged as shown in Table 23. The first entries describe the amount of data in TOTDAT. They are the number of physical and performance characteristics, NUMV, and the number of procurements, NUMS.

The next entry is an output designator called OUT. The value chosen will dictate the output option for the run. The options, together with the applicable value of OUT, will be described under data output.

The next entries describe the time groupings of the data in TOTDAT. The first entry, NUMT, defines the number of time groupings. It is followed by NUMT numbers (stored in a vector called ITBAT) which tell how many procurements are in each grouping. The effect of these numbers

TABLE 22

TEST RUN DATA

```
TOTDAT      13:19      LA "T"  Ø4/2Ø/69

1Ø1  967,2Ø4,144,31
1Ø2  4418,2Ø1,144,82
1Ø3  2414,217,149,6Ø
1Ø4  1996,178,153,95
1Ø5  852,172,1Ø7,25
1Ø6  2Ø72,215,136,67
1Ø7  1Ø4Ø8,221,177,243
1Ø8  2643,258,16Ø,54
1Ø9  3786,211,172,112
11Ø  3335,280,2Ø3,1Ø6
111  6374,3Ø5,196,183
112  7Ø92,294,187,156
113  1Ø3Ø4,28Ø,167,177
12Ø  1
121  2,4,3,1,5,6,7,8,9,1Ø,11,12,13
```

```
RUNDAT      13:18      LA"T"  Ø4/2Ø/69

1ØØ  3,13,1,9
1Ø2  5,1,1,1,1,1,1,1,1
1Ø3  Ø,Ø,Ø,Ø,Ø,5,6,7,8,9,1Ø,11,12,
1Ø4  2,1,3
```

TABLE 23

RUNDAT DATA

Line Number	Data File Values
1	NUMV, NUMS, OUT, NUMT
2	Vector ITBAT (NUMT entries)
3	Vector NW (NUMS entries)
4	NIV, NIV numbers (characteristic identifiers)
5	Repeat of line 4 for new PER
.	.
.	.
.	.
4+K	Repeat of line 4 for final PER

NUMV:	number of characteristics to be considered. Must have every characteristic called for by the PERs.
NUMS:	number of procurements in sample
OUT:	output designator
NUMT:	number of time batches
ITBAT:	vector for time grouping observations
NW:	vector for final output weights
K:	number of PERs
NIV:	number of independent variables for a particular PER

is to tell the program how many new procurements to include in the next sample that is to be given to the evaluation procedure. If the sample previously used contained the first n_1 procurements (from TOTDAT) and the next number in ITBAT is n_2 , then the next sample to be processed will consist of the first n_1+n_2 procurements (from TOTDAT).

The next entries are elements of the vector NW. There is one entry for each procurement. These are the weights that will be assigned to each residual for the calculation of Average Proportional Error and the other summary measures of bias and skewness (see Sec. IV B). They can be integer weights as the computer will divide by their sum.

The final entries in RUNDAT describe the PER to be used. The first entry corresponds to the number of physical and performance characteristics, NIV, and is followed by NIV numbers identifying the specific characteristics. For example, 2, 1, 3 would indicate that the PER consists of two characteristics and they are numbers 1 and 3. These latter numbers will tell the program which columns of TOTDAT to consider.

Provision in the program has been made to evaluate more than one PER in each computer run. Each PER must have the line of data just discussed (i.e., NIV and NIV characteristic identifiers). This is the only additional data needed, provided that all the independent variables are included in TOTDAT.

Test run values for RUNDAT are given in Table 22. In line 100, NUMV = 3, NUMS = 13, OUT = 1, NUMT = 9. The time groupings (vector ITBAT) are given in line 102. The first subsample will be 5 with one data point being added for each subsequent subsample.

The third line of data contains the weights for each of the residuals. No weight is given to the first 5, as no prediction of them will be made. Weights for the remaining points are the subsample size from which the prediction was made. Thus sample point 6 will have a weight of 5, point 7 will have a weight of 6, and so forth.

The final line of output indicates that the PER has two independent variables and they are variables 1 and 3.

Most of the data are entered into the program in Step 1 (Fig. 5), Input Data Base (lines 12-18, Table 17). This includes all the data with the exception of NIV and the characteristic designators. Ordering of the data base, if necessary, takes place in lines 19-28 of Table 17. In addition, the option to print the sample data from TOTDAT has been provided in lines 30-62, Table 17. The form of this output can be seen in Table 20 [output (1)]. If there are more than three independent variables, their values will be printed under the values for X1, X2 and X3 (e.g., X4 and X7 would appear under X1, etc.).

NIV and the characteristic designators are read in Step 2, Fig. 5, Estimation Procedure Description (lines 351-367, Table 18). The Step 2 data define the particular PER to be used. New PERs are also defined in Step 2 at the end of a loop from Step 10.

There are no limits on the number of PERs that can be evaluated in a given run. There are, however, upper limits on the number of procurements, NUMS, and number of characteristics, NUMV. These are currently programmed at 42 and 6; however, there is a tradeoff between them. From what I have been able to gather about the MARK I G.E. Time Sharing System, for which Historical Simulation has been programmed, all admissible combinations of upper limit values for NUMS and NUMV, for which any of the possible PER specifications (combinations of any subset of the NUMV variables) can be run, are given in Table 24. The table is stopped at NUMV = 12 for the reason that NUMV = 13 would yield an NUMS = 14 and thus not all 13 variables could be used as NUMS > NUMV+1 in order to fit the curves with a finite variance estimate. No advantage would be gained over the case when NUMV = 12 and it is possible to compile a larger sample (i.e., value of NUMS).

TABLE 24

POSSIBLE UPPER LIMIT VALUES FOR HISTORICAL SIMULATION
USING LINEAR PER-LEAST SQUARES

If NUMV =	1	2	3	4	5	6	7	8
Then NUMS ≤	183	115	83	63	51	42	35	30
If NUMV =	9	10	11	12				
Then NUMS ≤	26	22	19	16				

C. CALCULATIONS AND PREPARATION FOR OUTPUT

The actual program calculations are initiated in Step 2, Estimation Procedure Description (Fig. 5). In addition to the PER specification, discussed in the last section, the minimum sample size is calculated in this step (line 359, Table 18). The minimum sample size required depends on the PER and the technique being tested. For Linear PER-Least Square procedures the minimum sample size equals the number of independent variables in the PER (NIV) plus two (i.e., one larger than the number of parameters being estimated including the constant), so that estimates of variance are not infinite.

The final task performed in Step 2 is to print out a description of the estimating procedure being used (lines 375-399, Table 18). The output block (2), in Table 20, is printed out for Linear PER-Least Square procedures. In addition to the name, "LINEAR PER-LEAST SQUARES," the PER description consisting of number of independent variables (NIV = 2 in the example) and the characteristic numbers (1 and 3 in the example) are displayed. This block of output is repeated for each PER evaluated in the run. If a second PER were evaluated in the test run, this block of output for the new PER would appear after output block (7) in Table 20.

The next operation performed by the computer is to Set Up the Usable Data Base (Step 3, Fig. 5). The data matrix entered in Step 1 for TOTDAT is reduced in size by excluding characteristics not included in the PER defined in Step 2 (lines 73-89, Table 17). For the test run (Table 20) characteristic 2 is excluded from the rest of the PER evaluation.

Control is now passed to Step 4a (Fig. 5), in which the Initial Historical Sample Setup takes place. In lines 101-133, Table 17, data groupings, defined by the vector ITBAT, are added until the sample size is greater than or equal to the minimum sample size defined in Step 2. There may be situations in which the analyst wishes to specify a larger minimum sample size than the one automatically calculated. This can be done by making the first entry in ITBAT (see Table 23) the size of the minimum sample desired.

In the test run, NIV = 2 (line 106, RUNDAT Data, Table 22) and the first entry in ITBAT was 5 (first entry, line 102, same table). If ITBAT(1) ≤ 4, then the first subsample would have been equal to the minimum sample size, NIV+2 = 4. With ITBAT(1) = 5, however, the first subsample size is 5 [first output block (3), Table 20]. Hence the sample given to the estimation procedure consists of the first five procurements of TOTDAT with values for characteristic 1, characteristic 3, and the actual cost for each procurement. The sample size obtained is printed out as the first data block (3), Table 20.

The final operation in Step 4a is a housekeeping chore. A loop is set up for the remaining samples (lines 137-257, Table 17). The loop initiates with the number of entries in ITBAT used up to achieve the minimum sample size and is entered as many times as there are entries left in ITBAT. In the test case, the number of time groupings in ITBAT, NUMT, equaled 9. The values were 5, 1, 1, 1, 1, 1, 1, 1, 1. One entry was used in setting up the initial sample. The loop will therefore go from 2 to 9, resulting in eight more samples.

The new samples are defined in Step 4b (Fig. 5), Set Up Next Historical Sample, as the loop is reentered (lines 143-181, Table 17). Observations are added to the sample being passed to the evaluation procedure by adding the next n observations from the data base (Step 1), n being defined by ITBAT. For the test run this process results in sample sizes of 6 through 13 (the total sample for TOTDAT). As each sample is set up its size is printed out [output block (3), Table 20] to indicate that the next iteration is being started.

The sample defined in Step 4a or 4b is now passed to Step 5 (Fig. 5), where the computer Uses the Technique to Calculate the Parameters of the PER. In the test run, the technique is least squares, and the following operations are accomplished.

- Calculate arithmetic means of sample characteristics and costs (lines 610-630, Table 19)
- Calculate sample cross product matrix, i.e.,

  $\sum_i (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$ (lines 650-680, Table 19)*
- Invert cross product matrix (line 702, Table 20)**

* This is analogous to the S matrix referred to in Appendix II. The calculation is different in that the characteristic data are centered. The difference in the matrices is to prevent round-off errors from occurring in the computer (see Ref. 6, page 144).

**The program uses a matrix inversion routine that can be obtained from Ref. 11, program 9.6.

On the first iteration the sample being worked on is of size 5 (in our test run). At each succeeding iteration the sample expands to obtain new PER parameter estimates. These parameter estimates (calculated in lines 705-720, Table 19) are passed to Step 6b (by use of the common package) and define the CER for that iteration. In addition, the arithmetic means (lines 610-630, Table 19) are retained for use in Step 8b (by the common package).

Control is now passed to Step 6a (Fig. 5), where the machine Calculates Any Desired Iteration Output Statistics Valid for the Technique. A minimum output for any technique would be the PER parameter values. For some techniques this may be all that is desired.

For the Linear PER-Least Squares example being considered, the following operations are performed:

- Using the CER defined in 6b (lines 420-448, Table 18), calculate Fit Data for the sample (lines 721-801, Table 19). This includes an estimate for each procurement in the sample (5 for the first iteration), the standard error of the estimate, and an unadjusted R^2 (square of the multiple correlation coefficient).
- Print out (if desired) for each procurement the actual cost, estimated cost, cost difference, and proportional cost differences. This is shown as output block (4) in Table 20 and executed in lines 728 and 749 of Table 19.
- Calculate the variance-covariance matrix for the parameters (lines 806-845, Table 19). Deliver through the common package to Step 8b for use.
- Use diagonal elements of variance-covariance matrix to calculate t-statistics for each of the parameters (line 853, Table 19).
- In lines 846-855, Table 19, print subsample statistics, data block (5) in Table 20. These include the standard error of the estimate, R^2, the parameter values, the t-statistics, and the degrees of freedom.

Control is now passed to Step 7 (Fig. 5) where, in line 201 of Table 17, it is decided whether a new subsample must be processed. If the entire sample has been used, then control is passed to Step 9 (line 281). If not, there are procurements in the data base (Step 3) which were not used in estimating the parameters. Control is passed to Step 8a where the machine Makes Predictions for these procurements and Stores Data for Summary Statistics. For each procurement the following steps occur (lines 209-265, Table 17):

- Predict the cost of the procurement using the CER in 6b and characteristics from the data base in Step 3.
- Record actual cost, cost difference (from predicted), and proportional cost difference.
- Calculate any special statistics (line 233) from the technique using 8b (described below).
- Print prediction statistics, if desired. This is output block (6) in Table 20.
- Check to see if procurements will be included in next sample. If they will be, store the values calculated above for the summary output, data block (7).

The special statistics referred to above are calculated in Step 8b, Fig. 5 (lines 580-598, Table 18), Calculate Predictive Statistics Peculiar to Technique. For the Linear PER-Least Square Procedures, STAT 1 is the value of the t-distribution for the difference between the actual and predicted procurement cost times the standard deviation of the process, σ . Space has been left for a STAT 2 which is not presently used. The arithmetic means of the characteristics, calculated in Step 5, and the parameter variance-covariance matrix, calculated in Step 6a, are used to calculate the t-statistic for each procurement, together with the procurement informarion from Step 8a listed below:

The predicted cost

The actual cost

The characteristic values

The mathematical equations used to calculate these statistics are similar to those used to calculate prediction intervals. Given a new procurement, with characteristics $(x_1, \ldots, x_n)$, an actual cost A, a sample with m observations, $(X_{i1}, X_{i2}, \ldots, X_{in})$, $i = 1, 2, \ldots, m$, and a PER which contains n independent variables $(X_1, \ldots, X_n)$, then[*]

$$\frac{P - A}{SEE_m \sqrt{1 + \frac{1}{m} + DS^{-1}D}} \qquad (34)$$

has the t-distribution with $m - (n+1)$ degrees of freedom. In the above expression we have the following definitions:

S^{-1} = n·n inverse of the covariance matrix of the sample values of $X_{ik} - \overline{X}_i$ and $X_{jk} - \overline{X}_j$

D = n-dimensional column vector of terms $x_i - \overline{X}_i$

$\overline{X}_i$ = the arithmetic mean of the sample values of X_i

x_i = value of *ith* independent variable for the new procurement

SEE_m = the standard error of the estimate for sample size m

P = predicted value for the new procurement.

The quantity calculated in the computer program is Eq. 34 times SEE_m. This is the output that is used to calculate the statistics discussed in Sec. IV C.

[*] Reference 5, page 20.

At this point (line 257, Table 17), control passes to Step 4b where the next historical sample is set up (as previously discussed). Steps 5-7 are then repeated for the new sample. This process is continued until the entire data base (Step 3) is used. Since there are then no procurements to predict, control is passed to Step 9 (line 277, Table 17), Calculate Summary Statistics.

One value of each of the prediction outputs (data block 6, Table 20) has been saved for each procurement predicted. In the test run this would exclude procurements 1-5, as they were included in all the subsamples and hence never predicted. The values retained are those generated by the largest subsample used in the prediction of the particular procurement. These data are printed out in output block 7, Table 20 (line 301, Table 17). In the test run the procurements printed out were 6-13. Object 6 was estimated with five procurements in the sample, object 7 with six in the sample, and so forth.

These data are also used to calculate summary statistics (line 289-327, Table 17). At present these include the average proportional error, a measure of bias, and a measure of skewness. The weights in NW(I) from RUNDAT are used in these calculations. In the test cases, weights equal to the subsample size are used. Thus procurement 6 receives a weight of 5, procurement 7 a weight of 6, and so forth. For a discussion of these calculations, see Sec. IV B.

Control now passes to the final step of the program, Step 10, where it is determined whether a new PER is to be evaluated. If not, the run ends. If there is a new PER, as given by a new value for NIV and new characteristic numbers (Table 23), control is passed to Step 2 for new PER definition. In the test run there was no new PER defined, so the program terminated.

D. PROGRAM OUTPUT

The preceding paragraphs have described the complete output available for the program. This output, Table 20, can be divided into two classes: those printed in steps listed under the Estimation Procedure (Fig. 5) and those printed in the Main Program. Output data blocks (1), (3), (6), and (7) fall into this latter category. These blocks can be printed out no matter what estimation procedure is being tested. Their form will not change with different estimating procedures.

Data blocks (2), (4), and (5) are printed out in steps listed under the estimation procedure. Their form and content will change depending on the estimation procedure (or class of procedures) being tested.

The output of some of the data blocks is optional. The output designator, called OUT, is used to tell the machine what output to print. The following options are available:

Values of	Output Options (Excluded Blocks Checked)						
	Independent of Estimating Procedures				Valid for Particular Estimating Procedures		
OUT	(1)	(3)	(6)	(7)	(2)	(4)	(5)
1							
2						X	
3			X			X	
4	X						
5	X					X	
6	X		X			X	

Deleting the output of data blocks when they are not needed saves machine time and reduces the complexity of the output. The amount of output is greatly reduced when OUT = 6, as can be seen by visualizing Table 20 without data blocks (1), (4), and (6).

It should be noted that in any run with more than one PER, and a value of OUT 1 to 3, the sample [data block (1)] will be printed out only one time. The number 3 will be added to the value of OUT to prevent useless repetition of data block (1), which does not change with new PERs in a given run.

This concludes the presentation of the program currently available for Linear PER-Least Square Procedures. It should again be pointed out that operations on the left side of Fig. 5 are not dependent on the estimating procedure being evaluated. As subroutines for other classes of estimating procedures are developed, such as log-linear PERs, they will be tied into the operations on the left, the main program.

The program described in this appendix is operational on the G.E. Time-Sharing Service, MARK I. It therefore can be run on any terminal having that service. This flexibility tends to be offset by the slowness of the output via the teletype. Thirty-one minutes of console time was required for the test run.

Another drawback of working with the time-sharing service is the space limitation. As can be seen in Table 24, the size of the possible data base is not large, although for the cost application it seems adequate. Additions to the program cannot be made, however, without seriously diminishing this data base.

No attempt has been made to clean up the program to obtain greater efficiency. It is to be expected that improvements in operation time and space can be made by making the program or its output more efficient.

An alternative to this approach is to convert the program to a non-time-sharing machine; the program has been written in FORTRAN which should make conversion reasonably easy. There would certainly be savings in terminal time, although turn-around time will probably be longer (i.e., overnight service).

These suggestions for program improvement have not been implemerted to date as the program without any changes is adequate for demonstrating the Historical Simulation procedure and this was the purpose for which it was written.

APPENDIX II

DISTRIBUTION OF HISTORICAL SIMULATION PREDICTORS AND RESIDUALS UNDER THE USUAL MULTIPLE LINEAR REGRESSION MODEL ASSUMPTIONS

In this appendix it is assumed that the usual multiple linear regression model holds. From this assumption the distributions of the predictions and residuals obtained from Historical Simulation can be develope^d^ The reader is cautioned to keep in mind that these results are only valid when the multiple linear regression model assumptions are valid.

This appendix is divided into five parts or sections: The first establishes the multiple linear regression model assumptions; the second develops the Historical Simulation procedure (in the required notation); the third derives the distribution of predictions and residuals when only one subsample is used; the fourth derives similar distributions from two subsamples; and the last section summarizes the results.

In general, the predictions (residuals) are normally distributed and correlated. There are, however, some residuals which have zero covariance and this fact, together with normalcy, implies that they are independent. These are the one-step residuals, i.e., those residuals obtained from making the prediction of the next data point in time.

Before describing the multiple linear regression assumptions, it will be useful to discuss some of the mechanics of the statistical operators E for expected value and M for covariance matrix.* An understanding of their use in .latrix operations is a prerequisite to the understanding of this appendix.

*For a more complete discussion see Ref. 12, Sec. 2.4.

The expected value operator E is the easier of the two. Let $\vec{U}$, $\vec{V}$, and $\vec{W}$ be random vectors; A and B nonrandom transformation matrices; and $\vec{C}$ a vector of constants. Then the following relationship defines a set of linear equations

$$\vec{U} = A\vec{V} + B\vec{W} + \vec{C}$$

In this situation the expected value of $\vec{U}$ is defined by

$$E(\vec{U}) = AE(\vec{V}) + BE(\vec{W}) + \vec{C}$$

The covariance operator is a little harder to understand. Notationally it will be used in the three ways defined below:

1. Let $\vec{U}$ be a random vector (column); then its covariance matrix is given by

$$M_{\vec{U}} = E[U-E(U)][U-E(U)]'$$

where

[ ]' stands for the transpose of [ ]

2. Let $\vec{U}$ and $\vec{V}$ be two random vectors; then the covariance matrix between $\vec{U}$ and $\vec{V}$ is given by

$$M_{\vec{U},\vec{V}} = E[U-E(U)][V-E(V)]'$$

Note: $M_{\vec{U},\vec{V}} = M'_{\vec{V},\vec{U}}$

3. Let $\vec{U}$ and $\vec{V}$ be two random vectors; then their joint covariance matrix is given by

$$M_{\begin{bmatrix}\vec{U}\\ \vec{V}\end{bmatrix}} = \begin{bmatrix} M_{\vec{U}} & M_{\vec{U},\vec{V}} \\ M_{\vec{V},\vec{U}} & M_{\vec{V}} \end{bmatrix}$$

From these definitions it can be shown that if $\vec{R}$, $\vec{U}$, $\vec{V}$, and $\vec{W}$ are random vectors, and A is a nonrandom transformation matrix, and if

$$\vec{U} = A\vec{V}$$

and

$$\vec{R} = \vec{V} - \vec{W}$$

we have

$$M_{\vec{U}} = A\,M_{\vec{V}}\,A'$$

and

$$M_{\vec{R}} = M_{\vec{V}} + M_{\vec{W}} - M_{\vec{V},\vec{W}} - M_{\vec{W},\vec{V}}$$

In addition, we have

$$M_{\begin{bmatrix}\vec{U}\\ \vec{V}\end{bmatrix},\ \vec{W}} = \begin{bmatrix} M_{\vec{U},\vec{W}} \\ M_{\vec{V},\vec{W}} \end{bmatrix}$$

and

$$M_{\begin{bmatrix}\vec{U}\\ \vec{V}\end{bmatrix},\ \begin{bmatrix}\vec{R}\\ \vec{W}\end{bmatrix}} + \begin{bmatrix} M_{\vec{U},\vec{R}} & M_{\vec{U},\vec{W}} \\ M_{\vec{V},\vec{R}} & M_{\vec{V},\vec{W}} \end{bmatrix}$$

With an understanding of these operations in hand, we are ready to set out the multiple linear regression model assumptions.

A. MULTIPLE LINEAR REGRESSION ASSUMPTIONS[*]

The assumed sample consists of N P+1-tuples given by $(y_i, x_{i1}, x_{i2}, \ldots, x_{iP})$ for $i = 1, 2, \ldots, N$. For the Historical Simulation application, the P+1-tuples have been time ordered.

[*] See pages 384-388 in Ref. 7 for a more complete discussion of these assumptions.

The usual multiple linear regression hypotheses are given below:

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \qquad (35)$$

where

$\vec{Y}$ is a $N \times 1$ column vector whose transpose is given by

$$\vec{Y}' = (y_1, y_2, \ldots, y_N) \qquad (36)$$

X is an $N \times (P+1)$ matrix given by

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1P} \\ 1 & x_{21} & x_{22} & \cdots & x_{2P} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix} \qquad (37)$$

$\vec{\beta}$ is a $(P+1) \times 1$ column vector given by

$$\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_P \end{bmatrix} \qquad (38)$$

and $\vec{\epsilon}$ is an $N \times 1$ column vector given by

$$\vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_N \end{bmatrix} \qquad (39)$$

The matrix X and the vector $\vec{\beta}$ are assumed to be nonrandom while $\vec{\varepsilon}$ has a multivariate normal distribution, with mean vector equal to zero, a constant variance σ^2, and zero covariances. Hence

$$\varepsilon \overset{d}{=} N(0, \sigma^2 I_N) \tag{40}$$

where

I_N is an N-dimensional identity matrix

(The above notation is shorthand for "ε is normally distributed with $E(\varepsilon) = 0$ and $M_\varepsilon = \sigma^2 I_N$.")

Note that the $\underline{i}$th row of the X matrix is made up of the $\underline{i}$th P+1-tuple of the sample with a 1 replacing y_i. The 1 is used as the multiplier of the constant term β_0 in the regression equation. Defining

$$\vec{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{iP}) \quad ; \quad i = 1, 2, \ldots, N \tag{41}$$

we have for $i = 1, 2, \ldots, N$

$$y_i = \varepsilon_i + \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} = \vec{x}_i' \vec{\beta} + \varepsilon_i \tag{42}$$

Hence $\vec{Y}$ is a linear combination of the normal random variables in ε, and therefore it follows that $\vec{Y}$ is also a normally distributed random vector. The distribution can be derived from Eq. 40 and is given by

$$\vec{Y} \overset{d}{=} N(X\vec{\beta}, \sigma^2 I_N) \tag{43}$$

i.e., $E(\vec{Y}) = X\vec{\beta}$, and

$$M_{\vec{Y}} = \sigma^2 I_N$$

B HISTORICAL SIMULATION PROCEDURE

The regression assumptions fit into Historical Simulation as described in the following paragraphs.

Let n_o be the minimum sample size for Historical Simulation. It is necessary that n_o be greater than or equal to the smallest sample size necessary to carry out a linear regression analysis. Hence, $n_o \geq P+1$. For any n , $n_o \leq n < N$, define the following partition of the X matrix (defined in Eq. 37) by

$$X = \begin{array}{c} \left[\begin{array}{c} X_1^{(n)} \\ \hdashline X_2^{(n)} \end{array}\right] \end{array} \begin{array}{l} n \text{ rows} \\ N-n \text{ rows} \end{array} \qquad (44)$$

Also partition $\vec{Y}$ (defined in Eq. 35) in a similar manner obtaining

$$\vec{Y} = \begin{array}{c} \left[\begin{array}{c} \vec{Y}_1^{(n)} \\ \hdashline \vec{Y}_2^{(n)} \end{array}\right] \end{array} \begin{array}{l} n \text{ entries} \\ N-n \text{ entries} \end{array} \qquad (45)$$

Note that for this partition we have that $\vec{Y}_1^{(n)}$ and $\vec{Y}_2^{(n)}$ are independent, a consequence of Eq. 43. Furthermore, the joint covariance matrix breaks up as follows.

$$M_{\vec{Y}} = \sigma^2 I_N = \begin{array}{cc|l} n \text{ cols.} & N-n \text{ cols.} & \\ \sigma^2 I_n & 0 & n \text{ rows} \\ 0 & \sigma^2 I_{N-n} & \end{array}$$

Since

$$M_{\vec{Y}} = M_{\left[\begin{array}{c} \vec{Y}_1^{(n)} \\ \vec{Y}_2^{(n)} \end{array}\right]} = \left[\begin{array}{cc} M_{\vec{Y}_1^{(n)}} & M_{\vec{Y}_1^{(n)},\vec{Y}_2^{(n)}} \\ M_{\vec{Y}_2^{(n)},\vec{Y}_1^{(n)}} & M_{\vec{Y}_2^{(n)}} \end{array}\right]$$

we have

$$\left.\begin{aligned} & M_{\vec{Y}_1^{(n)}} = \sigma^2 I_n \\ & M_{\vec{Y}_2^{(n)}} = \sigma^2 I_{N-n} \\ \text{and} \quad & M_{\vec{Y}_1^{(n)}, \vec{Y}_2^{(n)}} = M_{\vec{Y}_2^{(n)}, \vec{Y}_1^{(n)}} = 0 \end{aligned}\right\} \tag{46}$$

If time batches are ignored, the Historical Simulation procedure (for the multiple linear regression model) can be defined as follows: For each n, $n_o \leq n < N$

1. Make a least squares fit using

$$\vec{Y}_1^{(n)} \quad \text{and} \quad X_1^{(n)}$$

2. Obtain an estimating vector of $\vec{\beta}$. Denote this vector $\hat{\vec{\beta}}^{(n)}$.

3. Use the resulting fit to make predictions of the remaining $N-n$ data points. This can be denoted by

$$\hat{\vec{Y}}^{(n)} = \begin{bmatrix} \hat{y}_{n+1}^{(n)} \\ \hat{y}_{n+2}^{(n)} \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_N^{(n)} \end{bmatrix} = X_2^{(n)} \hat{\vec{\beta}}^{(n)} \tag{47}$$

where

$\hat{y}_{n+k}^{(n)}$ is the prediction of y_{n+k} using a sample of size n.

4. Calculate the residuals by

$$\vec{D}^{(n)} = \begin{bmatrix} d_{n+1}^{(n)} \\ d_{n+2}^{(n)} \\ \cdot \\ \cdot \\ \cdot \\ d_N^{(n)} \end{bmatrix} = \begin{bmatrix} \hat{y}_{n+1}^{(n)} - y_{n+1} \\ \hat{y}_{n+2}^{(n)} - y_{n+2} \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_N^{(n)} - y_N \end{bmatrix} = \hat{Y}^{(n)} - Y_2^{(n)} \tag{48}$$

where $d_{n+k}^{(n)}$ denotes the difference (residual) between the predicted $\hat{y}_{n+k}^{(n)}$ and the actual y_{n+k} .

After the Historical Simulation procedure is completed,

$$\sum_{n=n_o}^{N-1} (N-n)$$

predictions and the same number of residuals will have been calculated. These are denoted by the random vectors

$$\left[\vec{\hat{Y}}^{(n_o)}, \vec{\hat{Y}}^{(n_o+1)}, \ldots, \vec{\hat{Y}}^{(N-1)}\right]$$

and

$$\left[D^{(n_o)}, D^{(n_o+1)}, \ldots, D^{(N-1)}\right]$$

respectively. The problem is to find the distribution of these random vectors.

C. DISTRIBUTION OF $\vec{\hat{Y}}^{(n)}$ AND $\vec{D}^{(n)}$, i.e., THE PREDICTIONS AND THE RESIDUALS, FROM ONE SAMPLE SIZE n

From Eq. 47 it can be seen that the form and distribution of $\vec{\hat{B}}^{(n)}$ must be established before the distribution of $\vec{\hat{Y}}^{(n)}$ and $\vec{D}^{(n)}$ can be developed. The distribution of $\hat{\vec{\beta}}^{(n)}$ is well known,[*] and hence only the results and a sketch of the reasoning will be presented here. Let

$$S^{(n)} = X_1'^{(n)} X_1^{(n)} \tag{49}$$

where $X_1'^{(n)}$ is the transpose of $X_1^{(n)}$. Then the least squares solution for the parameter vector $\vec{\beta}$ is given by

$$\hat{\vec{\beta}}^{(n)} = S^{(n)^{-1}} X_1'^{(n)} \vec{Y}_1^{(n)} \tag{50}$$

where $S^{(n)^{-1}}$ is the inverse of $S^{(n)}$.

Now, the only random variables in Eq. 50 are $\vec{Y}_1^{(n)}$. Hence $\hat{\vec{\beta}}^{(n)}$ is a linear combination of the normal random variables $y_1, y_2, \ldots, y_n$ and hence are normally distributed. The expected value and the variance-covariance matrix are given as follows.

$$E\hat{\vec{\beta}}^{(n)} = \vec{\beta}$$

and

$$M_{\hat{\vec{\beta}}^{(n)}} = \sigma^2 S^{(n)^{-1}}$$

Hence, we have that

$$\hat{\vec{\beta}}^{(n)} \overset{d}{=} N(\vec{\beta}, \sigma^2 S^{(n)^{-1}}) \tag{51}$$

Furthermore, since $S^{(n)^{-1}}$ is essentially a covariance matrix, it is symmetric; therefore

$$S^{(n)^{-1}} = S'^{(n)^{-1}} \tag{52}$$

Now, from Eq. 47 the predictions $\hat{\vec{Y}}^{(n)}$ are given by $\hat{\vec{Y}}^{(n)} = X_2^{(n)} \hat{\vec{\beta}}^{(n)}$. Hence, $\hat{\vec{Y}}^{(n)}$ are linear combinations of the normal random variables

[*] For further details see p. 386 in Ref. 7.

$\hat{\vec{\beta}}^{(n)}$ and are therefore normally distributed. Th' mean vector and covariance matrix are calculated below.

$$E\hat{\vec{Y}}^{(n)} = X_2^{(n)} E\,\hat{\vec{\beta}}^{(n)}$$

$$= X_2^{(n)}\vec{\beta} \qquad \text{from Eq. 51}$$

$$= E\vec{Y}_2^{(n)} \qquad \text{from Eqs. 43, 44, and 45}$$

$$M_{\hat{\vec{Y}}^{(n)}} = X_2^{(n)} M_{\hat{\vec{\beta}}^{(n)}} X_2'^{(n)}$$

$$= \sigma^2 X_2^{(n)} S^{(n)^{-1}} X_2'^{(n)} \qquad \text{from Eq. 51}$$

Hence, we have that

$$\hat{\vec{Y}}^{(n)} \overset{d}{=} N\left[E\vec{Y}_2^{(n)}\ ,\ M_{\hat{\vec{Y}}^{(n)}}\right] \qquad (53)$$

where

$$E\hat{\vec{Y}}^{(n)} = X_2^{(n)}\vec{\beta} = E\vec{Y}_2^{(n)}$$

and

$$M_{\hat{\vec{Y}}^{(n)}} = \sigma^2 X_2^{(n)} S^{(n)^{-1}} X_2'^{(n)}$$

Finally, the distribution of the residuals $\vec{D}^{(n)}$ can now be found. From Eq. 48 it will be recalled that

$$\vec{D}^{(n)} = \hat{\vec{Y}}^{(n)} - \vec{Y}_2^{(n)}$$

which is the difference of two normally distributed random vectors. Hence, $\vec{D}^{(n)}$ is normal. The mean vector and covariance matrix are calculated below.

$$E\left[\vec{D}^{(n)}\right] = E\left[\hat{\vec{Y}}^{(n)}\right] - E\left[\vec{Y}_2^{(n)}\right]$$

$$= E\left[\vec{Y}_2^{(n)}\right] - E\left[\vec{Y}_2^{(n)}\right] \qquad \text{from Eq. 53}$$

$$= 0$$

and

$$M_{\vec{D}^{(n)}} = M_{\hat{\vec{Y}}^{(n)}} + M_{\vec{Y}_2^{(n)}} - M_{\vec{Y}_2^{(n)}, \hat{\vec{Y}}^{(n)}} - M_{\hat{\vec{Y}}^{(n)}, \vec{Y}_2^{(n)}} \tag{54}$$

Now

$$\hat{\vec{Y}}^{(n)} = X_2^{(n)} S^{(n)^{-1}} X_1'^{(n)} \vec{Y}_1^{(n)} \qquad \text{from Eqs. 47 and 48}$$

Hence

$$M_{\hat{\vec{Y}}^{(n)}, \vec{Y}_2^{(n)}} = X_2^{(n)} S^{(n)^{-1}} X_1'^{(n)} M_{\vec{Y}_1^{(n)}, \vec{Y}_2^{(n)}}$$

But from Eq. 46 we have that

$$M_{\vec{Y}_1^{(n)}, \vec{Y}_2^{(n)}} = 0$$

as $\vec{Y}_1^{(n)}$ and $\vec{Y}_2^{(n)}$ are independent. Hence,

$$M_{\hat{\vec{Y}}^{(n)}, \vec{Y}_2^{(n)}} = 0$$

By similar reasoning it can be shown that

$$M_{\vec{Y}_2^{(n)},\,\vec{Y}^{(n)}} = 0$$

Hence we have from Eq. 54 that

$$M_{\vec{D}^{(n)}} = M_{\vec{Y}^{(n)}} + M_{\vec{Y}_2^{(n)}}$$

$$= \sigma^2 \left[X_2^{(n)} S^{(n)^{-1}} X_2'^{(n)} + I_{N-n} \right]$$

from Eqs. 43 and 53

To summarize, then, we have that

$$\vec{D}^{(n)} \overset{d}{=} N \left[0, M_{\vec{D}^{(n)}} \right] \qquad (55)$$

where

$$M_{\vec{D}^{(n)}} = M_{\vec{Y}^{(n)}} + M_{\vec{Y}_2^{(n)}} = \sigma^2 \left[X_2^{(n)} S^{(n)^{-1}} X_2'^{(n)} + I_{N-n} \right]$$

D. DISTRIBUTION OF PREDICTIONS AND RESIDUALS FROM TWO SUBSAMPLES

So far we have been addressing the distribution of the predictions and the residuals from one subsample. The problem now is to find the joint distributions of

$$\hat{\vec{Y}}^{(n_1)},\ \hat{\vec{Y}}^{(n_2)}$$

and the joint distribution of

$$\vec{D}^{(n_1)},\ \vec{D}^{(n_2)}$$

where $n_o \leq n_1 < n_2 < N$. That they are normal has been determined in Sec. C, as has the value of their expected values. In addition, we have determined part of their covariance matrices in Sec. C, namely

$$M_{\vec{Y}^{(n_1)}} \quad \text{and} \quad M_{\vec{Y}^{(n_2)}} \qquad\qquad \text{from Eq. 53}$$

and

$$M_{\vec{D}^{(n_1)}} \quad \text{and} \quad M_{\vec{D}^{(n_2)}} \qquad\qquad \text{from Eq. 55}$$

Still to be determined are $M_{\vec{Y}^{(n_1)},\vec{Y}^{(n_2)}}$ and $M_{\vec{D}^{(n_1)},\vec{D}^{(n_2)}}$ (and their transposes).

At this point it is necessary to introduce some additional notation. Let the X matrix and the $\vec{Y}$ vector be partitioned as follows:

$$X = \begin{array}{c} \left[\begin{array}{c} X_1^{(n_1)} \\ \\ \dot{X} \\ \\ X_2^{(n_2)} \end{array} \right] \end{array} \begin{array}{l} n_1 \text{ rows} \\ \\ n_2-n_1 \text{ rows} \\ \\ N-n_2 \text{ rows} \end{array}$$

and $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (56)

$$\vec{Y} = \begin{array}{c} \left[\begin{array}{c} \vec{Y}_1^{(n_1)} \\ \\ \dot{Y} \\ \\ \vec{Y}_2^{(n_2)} \end{array} \right] \end{array} \begin{array}{l} n_1 \text{ entries} \\ \\ n_2-n_1 \text{ entries} \\ \\ N-n_2 \text{ entries} \end{array}$$

The relationship of this new partition to the partition previously discussed is as follows:

$$\left.\begin{aligned}
&\left[\begin{array}{c} X_1^{(n_1)} \\ \hdashline X_2^{(n_1)} \end{array}\right] = \left[\begin{array}{c} X_1^{(n_1)} \\ \hdashline \dot{X} \\ \hdashline X_2^{(n_2)} \end{array}\right] = \left[\begin{array}{c} X_1^{(n_2)} \\ \hdashline X_2^{(n_2)} \end{array}\right] \\
\text{and} \\
&\left[\begin{array}{c} \vec{Y}_1^{(n_1)} \\ \hdashline \vec{Y}_2^{(n_1)} \end{array}\right] = \left[\begin{array}{c} \vec{Y}_1^{(n_1)} \\ \hdashline \dot{Y} \\ \hdashline \vec{Y}_2^{(n_2)} \end{array}\right] = \left[\begin{array}{c} \vec{Y}_1^{(n_2)} \\ \hdashline \vec{Y}_2^{(n_2)} \end{array}\right]
\end{aligned}\right\} \qquad (57)$$

Hence $\dot{X}$ and $\dot{Y}$ are the regions of shift in the partition when the Historical Simulation procedure passes from a subsample of size n_1 to one of size n_2. $\dot{X}$ is not used in the fit routine when the subsample size is n_1, but it is used when the subsample size is n_2. Similarly, predictions are made of $\dot{Y}$ when the subsample size is n_1, but they are not made when the subsample size is n_2.

By analogous reasoning to that used to establish Eq. 46 and the mutual independence of $\vec{Y}_1^{(n_1)}$, $\dot{Y}$, $\vec{Y}_2^{(n_2)}$, which is a consequence of Eq. 43, it can be shown that

$$\left.\begin{aligned}
M_{\dot{Y}_1(n_1)} &= \sigma^2 I_{n_1} \\
M_{\dot{Y}} &= \sigma^2 I_{n_2-n_1} \\
M_{\dot{Y}(n_2)} &= \sigma^2 I_{N-n_2} \\
\text{and} \quad M_{\dot{Y}_1^{(n_1)}, \dot{Y}}, \; M_{\dot{Y}_1^{(n_1)}, \dot{Y}_2^{(n_2)}}, \; M_{\dot{Y}, \dot{Y}_2(n_2)}
\end{aligned}\right\} \quad (58)$$

[and their transposes all equal zero.]

We now turn back to the problem of calculating

$$M_{\hat{Y}^{(n_1)}, \hat{Y}^{(n_2)}} \quad \text{and} \quad M_{\dot{D}^{(n_1)}, \dot{D}^{(n_2)}}$$

These are calculated below:

From (47) and (50) we have that

$$\hat{Y}^{(n_1)} = X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} \dot{Y}_1^{(n_1)}$$

and
$$\hat{Y}^{(n_2)} = X_2^{(n_2)} S^{(n_2)^{-1}} X_1'^{(n_2)} \dot{Y}_1^{(n_2)}$$

Hence

$$M_{\hat{Y}^{(n_1)}, \hat{Y}^{(n_2)}} = X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} M_{\dot{Y}_1^{(n_1)}, \dot{Y}_1^{(n_2)}} X_1^{(n_2)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

But

$$\dot{Y}_1'^{(n_2)} = \left[\dot{Y}_1'^{(n_1)} \;\middle|\; \dot{Y}' \right]$$

Hence

$$M_{\dot{Y}_1^{(n_1)}, \dot{Y}_1^{(n_2)}} = \left[M_{\dot{Y}_1^{(n_1)}} M_{\dot{Y}^{(n_1)}, \dot{Y}} \right]$$

$$= \left[\sigma^2 I_{n_1} \;\middle|\; 0 \right] \qquad \text{by Eq. 58}$$

Hence we have

$$M_{\hat{Y}^{(n_1)}, \hat{Y}^{(n_2)}} = X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} \left[\sigma^2 I_{n_1} \;\middle|\; 0 \right] X_1^{(n_2)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

But

$$X_1^{(n_2)} = \begin{bmatrix} X_1^{(n_1)} \\ \\ \dot{X} \end{bmatrix} \qquad \text{from Eq. 57}$$

Hence

$$\left[\sigma^2 I_{n_1} \;\middle|\; 0 \right] X_1^{(n_2)} = \sigma^2 I_{n_1} X_1^{(n_1)} + 0 = \sigma^2 X_1^{(n_1)}$$

By substitution we then have

$$M_{\hat{Y}^{(n_1)}, \hat{Y}^{(n_2)}} = \sigma^2 X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} X_1^{(n_1)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

$$= \sigma^2 X_2^{(n_1)} S^{(n_1)^{-1}} S^{(n_1)} S'^{(n_2)^{-1}} X_2'^{(n_2)} \qquad \text{from Eq. 49}$$

$$= \sigma^2 X_2^{(n_1)} S^{(n_2)^{-1}} X_2'^{(n_2)} \qquad \text{from Eq. 52}$$

$$= \sigma^2 \begin{bmatrix} \dot{X} \\ \\ X_2^{(n_2)} \end{bmatrix} S^{(n_2)^{-1}} X_2'^{(n_2)} \qquad \text{from Eq. 57}$$

But from Eq. 53 we have

$$M_{\vec{Y}^{(n_2)}} = \sigma^2 X_2^{(n_2)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

Hence

$$M_{\vec{Y}^{(n_1)},\vec{Y}^{(n_2)}} = \begin{bmatrix} \sigma^2 XS^{(n_2)^{-1}} X_2'^{(n_2)} \\ \\ M_{\vec{Y}^{(n_2)}} \end{bmatrix} \qquad (59)$$

The joint distribution of $\vec{Y}^{(n_1)},\vec{Y}^{(n_2)}$ can now be summarized from Eqs. 53 and 59 as follows:

$$\begin{bmatrix} \vec{Y}^{(n_1)} \\ \\ \vec{Y}^{(n_2)} \end{bmatrix} \overset{d}{=} N\left[\begin{bmatrix} EY_2^{(n_1)} \\ \\ EY_2^{(n_2)} \end{bmatrix}, M_{\begin{bmatrix} \vec{Y}^{(n_1)} \\ \vec{Y}^{(n_2)} \end{bmatrix}}\right] \qquad (60)$$

where

$$M_{\begin{bmatrix} \vec{Y}^{(n_1)} \\ \vec{Y}^{(n_2)} \end{bmatrix}} = \begin{array}{cc|c|l} \multicolumn{2}{c}{N-n_1 \text{ Cols}} & N-n_2 \text{ Cols} & \\ \hline \multicolumn{2}{c|}{} & A & n_2-n_1 \text{ rows} \\ \cline{3-3} \multicolumn{2}{c|}{M_{\vec{Y}^{(n_1)}}} & M_{\vec{Y}^{(n2)}} & N-n_2 \text{ rows} \\ \hline A & \multicolumn{1}{|c|}{M_{\vec{Y}^{(n_2)}}} & M_{\vec{Y}^{(n_2)}} & N-n_2 \text{ rows} \\ \hline n_2-n_1 \text{ Cols} & N-n_2 \text{ Cols} & & \end{array}$$

$$A = \sigma^2 XS^{(n_2)^{-1}} X_2'^{(n_2)}$$

and the other entries are defined by Eq. 53.

It is rather interesting to note that the covariances defined in (59) are not dependent on

$$S^{(n_1)^{-1}}$$

The implications of this are that the covariance of predictions made from different subsample sizes are dependent only on the data matrix used in the larger subsample size fit. In particular

$$\mathrm{COV}\left[\hat{y}^{(n_1)}_{n_1+k}\ ,\ \hat{y}^{(n_2)}_{n_2+j}\right] = \mathrm{COV}\left[\hat{y}^{(n_2)}_{n_1+k}\ ,\ \hat{Y}^{(n_2)}_{n_2+j}\right] \tag{61}$$

for $n_1+k > n_2$; i.e., y_{n_1+k} was predicted from both subsamples.

Turning to the last task,

$$M_{\vec{D}^{(n_1)},\vec{D}^{(n_2)}}$$

must be calculated. From Eq. 48 we have that

$$\vec{D}^{(n_1)} = \hat{\vec{Y}}^{(n_1)} - \vec{Y}_2^{(n_1)}$$

and

$$\vec{D}^{(n_2)} = \hat{\vec{Y}}^{(n_2)} - \hat{\vec{Y}}_2^{(n_2)}$$

Hence

$$M_{\vec{D}^{(n_1)},\vec{D}^{(n_2)}} = M_{\hat{\vec{Y}}^{(n_1)},\hat{\vec{Y}}^{(n_2)}} + M_{\vec{Y}_2^{(n_1)},\vec{Y}_2^{(n_2)}} - M_{\vec{Y}_2^{(n_1)},\hat{\vec{Y}}^{(n_2)}}$$

$$- M_{\hat{\vec{Y}}^{(n_1)},\vec{Y}_2^{(n_2)}}$$

Taking this expression term by term we have that

$$M_{\hat{\vec{Y}}^{(n_1)},\hat{\vec{Y}}^{(n_2)}} = \left[\begin{array}{c} \sigma^2 \dot{X} S^{(n_2)^{-1}} X_2'^{(n_2)} \\ \hdashline M_{\vec{Y}^{(n_2)}} \end{array}\right] \qquad \text{by Eq. 59}$$

$$M_{\vec{Y}_2^{(n_1)},\vec{Y}_2^{(n_2)}} = M_{\begin{bmatrix} \dot{Y} & \\ & \vec{Y}_2^{(n_2)} \\ \vec{Y}^{(n_2)} & \end{bmatrix}} = \begin{bmatrix} M_{\dot{Y},\vec{Y}_2^{(n_2)}} \\ \\ M_{\vec{Y}_2^{(n_2)}} \end{bmatrix} \qquad \text{from Eq. 57}$$

$$= \left[\begin{array}{c} 0 \\ \hdashline \sigma^2 I_{N-n_2} \end{array}\right] \qquad \text{from Eq. 58}$$

Hence

$$M_{\vec{Y}_2^{(n_1)},\vec{Y}_2^{(n_2)}} = \left[\begin{array}{c} 0 \\ \hdashline \sigma^2 I_{N-n_2} \end{array}\right]$$

From Eqs. 47 and 50 we have that

$$\hat{\vec{Y}}^{(n_2)} = X_2^{(n_2)} S^{(n_2)^{-1}} X_1'^{(n_2)} \vec{Y}_1^{(n_2)}$$

Hence

$$M_{\vec{Y}_2^{(n_1)},\hat{\vec{Y}}^{(n_2)}} = M_{\vec{Y}_2^{(n_1)},\vec{Y}_1^{(n_2)}} X_1^{(n_2)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

But by Eq. 57

$$\vec{Y}_2^{(n_1)} = \begin{bmatrix} \dot{Y} \\ \\ \vec{Y}_2^{(n_2)} \end{bmatrix} \quad \text{and} \quad \vec{Y}_1^{(n_2)} = \begin{bmatrix} \vec{Y}_1^{(n_1)} \\ \\ \dot{Y} \end{bmatrix}$$

Hence

$$M_{\vec{Y}_2^{(n_1)},\vec{Y}_1^{(n_2)}} = \left[\begin{array}{c|c} M_{\dot{Y},\vec{Y}_1^{(n_1)}} & M_{\dot{Y},\dot{Y}} \\ \hline M_{\vec{Y}_2^{(n_2)},\vec{Y}_1^{(n_1)}} & M_{\vec{Y}_2^{(n_2)},\dot{Y}} \end{array}\right] = \left[\begin{array}{c|c} 0 & \sigma^2 I_{n_2-n_1} \\ \hline 0 & 0 \end{array}\right]$$

by Eq. 58

Hence

$$M_{\vec{Y}_2^{(n_1)},\vec{Y}^{(n_2)}} = \left[\begin{array}{c|c} 0 & \sigma^2 I_{n_2-n_1} \\ \hline 0 & 0 \end{array}\right] X_1^{(n_2)} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

But

$$X_1^{(n_2)} = \begin{bmatrix} X_1^{(n_1)} \\ \\ \dot{X} \end{bmatrix}$$

by Eq. 57

Hence

$$M_{\vec{Y}^{(n_1)},\vec{Y}^{(n_2)}} = \begin{bmatrix} 0 & \sigma^2 I_{n_2-n_1} \\ \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1^{(n_1)} \\ \\ \dot{X} \end{bmatrix} S^{(n_2)^{-1}} X_2'^{(n_2)}$$

Performing the multiplication we then have

$$M_{\vec{Y}_2^{(n_1)},\vec{Y}^{(n_2)}} = \left[\begin{array}{c} \sigma^2 \dot{X} S^{(n_2)^{-1}} X_2'^{(n_2)} \\ \hline 0 \end{array}\right]$$

The final term is

$$M_{\vec{Y}^{(n_1)},\vec{Y}_2^{(n_2)}}$$

From Eqs. 47 and 50 we have that

$$\hat{\vec{Y}}^{(n_1)} = X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} \vec{Y}_1^{(n_1)}$$

Hence

$$M_{\hat{\vec{Y}}_1^{(n_1)}, \vec{Y}_2^{(n_2)}} = X_2^{(n_1)} S^{(n_1)^{-1}} X_1'^{(n_1)} M_{\vec{Y}_1^{(n_1)}, \vec{Y}_2^{(n_2)}}$$

But

$$M_{\vec{Y}_1^{(n_1)}, \vec{Y}_2^{(n_2)}} = 0 \qquad \text{from Eq. 58}$$

Hence

$$\underline{M_{\hat{\vec{Y}}^{(n_1)}, \vec{Y}_2^{(n_2)}} = 0}$$

Collecting underlined terms we therefore have that

$$M_{\vec{D}^{(n_1)}, \vec{D}^{(n_2)}} = \left[\begin{array}{c} \sigma^2 \dot{X} S^{(n_2)^{-1}} X_2'^{(n_2)} \\ \hdashline M_{\hat{\vec{Y}}^{(n_2)}} \end{array}\right] + \left[\begin{array}{c} 0 \\ \hdashline \sigma^2 I_{N-n_2} \end{array}\right]$$

$$- \left[\begin{array}{c} \sigma^2 \dot{X} S^{(n_2)^{-1}} X_2'^{(n_2)} \\ \hdashline 0 \end{array}\right] - 0$$

Hence

$$M_{\vec{D}^{(n_1)}, \vec{D}^{(n_2)}} = \left[\begin{array}{c} 0 \\ \hdashline M_{\vec{Y}^{(n_2)}} + \sigma^2 I_{N-n_2} \end{array}\right]$$

and by Eq. 55 we have

$$M_{\vec{D}^{(n_1)},\vec{D}^{(n_2)}} = \begin{bmatrix} 0 \\ ---- \\ M_{\vec{D}^{(n_2)}} \end{bmatrix} \tag{62}$$

This result has some rather interesting consequences. As in the case of the predictions, covariance between residuals depends only on the information used in the fit performed on the larger of the two subsamples used to generate the residuals. Hence, we have an analogous result to Eq. 60, namely

$$\mathrm{COV}\left[d^{(n_1)}_{n_1+k}\,,\ d^{(n_2)}_{n_2+j}\right] = \mathrm{COV}\left[d^{(n_2)}_{n_1+k}\,,\ d^{(n_2)}_{n_2+j}\right] \tag{63}$$

for $n_1+k > n_2$; i.e., residual calculations were made for y_{n_1+k} from both subsamples.

An even more interesting consequence is that the covariance between residuals, one of which is not calculated for both subsamples, is zero. Hence

$$\mathrm{COV}\left[d^{(n_1)}_{n_1+k}\,,\ d^{(n_2)}_{n_2+j}\right] = 0 \tag{64}$$

if $n_1+k \leq n_2$. In particular, the one-step residuals

$$\left[d^{(n_o)}_{n_o+1}\,, \ldots\ldots,\ d^{(n)}_{n+1}\,, \ldots\ldots,\ d^{(N-1)}_{N}\right]$$

have zero covariances. This fact, coupled with a normal distribution, implies independence.* Hence, the one-step residuals are mutually independent.

---

*The proof of this assertion is not too difficult. Consider two normally distributed random variables U and V . Let

$$U \overset{d}{=} N\left[\, E(U), \sigma_U^2 \,\right]$$

$$V \overset{d}{=} N\left[\, E(V), \sigma_V^2 \,\right]$$

and

$$COV(UV) = 0$$

Then

$$f_U(u) = \frac{1}{\sqrt{2\pi\ \sigma_U^2}} \ \exp - \left[\frac{(u - EU)^2}{2\sigma_U^2} \right]$$

$$f_V(v) = \frac{1}{\sqrt{2\pi\ \sigma_V^2}} \ \exp - \left[\frac{(v - EV)^2}{2\sigma_V^2} \right]$$

and

$$f_{U,V}(u,v) = \frac{1}{2\pi\sigma_U\sigma_V} \ \exp - \left[\frac{(u - EU)^2}{2\sigma_U^2} + \frac{(v - EV)^2}{2\sigma_V^2} \right]$$

where f_U and f_V are the density functions of U and V and $f_{U,V}$ is the joint density function of U and V .

Now, according to Ref. 7, page 131, U and V are independent if

$$f_{U,V} = (f_U)\ (f_V)$$

This is clearly the case for the densities defined above.

The joint distribution of $\vec{D}^{(n_1)}$ and $\vec{D}^{(n_2)}$ can now be summarized using Eqs. 55 and 62. We have that

$$\begin{bmatrix}\vec{D}^{(n_1)}\\ \vec{D}^{(n_2)}\end{bmatrix} \overset{d}{=} N\left[0\ ,\ M_{\begin{bmatrix}\vec{D}^{(n_1)}\\ \vec{D}^{(n_2)}\end{bmatrix}}\right] \tag{65}$$

where

$$M_{\begin{bmatrix}\vec{D}^{(n_1)}\\ \vec{D}^{(n_2)}\end{bmatrix}} = \begin{array}{cc|cl}
\multicolumn{2}{c|}{N-n_1\text{ Cols}} & N-n_2\text{ Cols} & \\
\hline
\multicolumn{2}{c|}{M_{\vec{D}^{(n_1)}}} & 0 & n_2-n_1\text{ rows}\\
\cline{3-3}
\multicolumn{2}{c|}{} & M_{\vec{D}^{(n_2)}} & N-n_2\text{ rows}\\
\hline
0 & M_{\vec{D}^{(n_2)}} & M_{\vec{D}^{(n_2)}} & N-n_2\text{ rows}\\
\hline
n_2-n_1\text{ Cols} & N-n_2\text{ Cols} & N-n_2\text{ Cols} &
\end{array}$$

E. SUMMARY

With the help of some additional notation, it is possible to summarize and simplify the results of the preceding sections and combine all the information about the form and distribution of the predictions and residuals into one table. Recall from Eq. 41 that $\vec{x}_i'$ is a row vector equal to the _i_th row of the X matrix. Now, define a set of constants by

$$C_j^{n+i} = \vec{x}_{n+i}' S^{(n)^{-1}} \vec{x}_j$$

where $\vec{x}_j$ is a column vector corresponding to the _j_th column of X' (or the jth row of X).

Then by examining the results of sections B, C, and D of this appendix, one can obtain the form and distribution of the Predictions and Residuals given in Table 25. Reference numbers of the equations and facts derived in sections B, C, and D are given in parentheses.

Some of the conclusions that can be drawn from this table are given below.

1. The residual covariances are very similar to the prediction covariances. In fact, they are equal unless one of the points being predicted is not predicted from both subsamples, or the point being predicted is the same for both subsamples. In the first of these exceptions the residual covariance is zero.

In the second exception, the calculation is similar to a variance calculation. The residual variance is obtained by adding σ^2 to the prediction variance. In like manner, the residual covariance is obtained by adding σ^2 to the prediction covariance.

2. The covariances depend on the two subsample sizes m and n only insofar as which S-matrix to use. If $m \leq n$ then $S^{(n)}$ is used. If $m > n$ then $S^{(m)}$ is used. This is the only difference between the coefficients C_{n+k}^{m+j} and C_{m+j}^{n+k}. The rule to follow is always use the S-matrix corresponding to the larger sample size.

3. In general the predictions and residuals are correlated (among themselves). However,

$$\mathrm{COV}\left[d_{n+k}^{(n)}, \dot{d}_{m+j}^{(m)}\right] = 0$$

if $n+k \leq m$ or $m+j \leq n$. This means in particular that the one-step residuals $d_{n+1}^{(n)}$, $n_0 \leq n < N$, are uncorrelated. As discussed in section D, this zero covariance, together with a normal distribution, implies independence.

TABLE 25

FORM AND DISTRIBUTION OF PREDICTIONS AND RESIDUALS

(Assuming the usual multiple linear regression model assumptions)

Notation	Prediction $y_{n+k}^{(n)}$		Residual $d_{n+k}^{(n)}$	
Sample point for which the prediction (residual) pertains	$n+k \;;\; 0 < k \leq N-n$		$n+k \;;\; 0 < k \leq N-n$	
Subsample size used	n		n	
Calculation	$\sum_{j=1}^{n} c_j^{(n+k)} y_j$	(47) & (50)	$y_{n+k}^{(n)} - y_{n+k}$	(48)
Distribution	Normal	(53)	Normal	(55)
Expected value	$x'_{n+k}\beta$	(53)	0	(55)
Variance	$\sigma^2 c_{n+k,n+k}$	(53)	$\sigma^2(1 + c_{n+k,n+k})$	(55)
Covariance with	$y_{m+j}^{(m)}$		$d_{m+j}^{(m)}$	
[illegible]	[illegible]		[illegible]	[illegible]
[illegible]	$\sigma^2 c_{n+k,m+j}$	(53) & (55)	$\sigma^2 c_{n+k,m+j}$	(55)
[illegible]			$\sigma^2(1 + c_{n+k,m+j})$	[illegible]

([illegible] with $c_{m+j,n+k}$ replacing $c_{n+k,m+j}$)

[illegible] matrix [illegible] defined in [illegible]

[illegible]

[illegible] parameters defined in [illegible]

[illegible] parameters defined in [illegible]

[illegible]

APPENDIX III

PROPERTIES OF VARIANCE ESTIMATORS

In this appendix, as in Appendix II, the multiple linear regression model will be assumed. This model has been described in Appendix II, Eqs. 35 through 43.

In Sec. IV C of the body of the report, modified residuals were derived which are theoretically a random sample from a normal population with mean 0 and variance σ^2--the variance of the error terms ϵ in the regression model, Eq. 40. These modified residuals are denoted by $r = (r_{n_o}, r_{n_o+1}, \ldots, r_{N-1})$ where n_o is the minimum sample size for Historical Simulation and N is the size of the data base.

A Kolmogorov-Smirnov (K-S) Goodness-of-Fit Test has been suggested, in Sec. IV C. In order to apply this test, however, an estimate of the variance σ^2 must be made. Three possible candidates have been considered. In this appendix these candidates are described; distributions are derived for the case when the multiple linear regression model holds (which is the null hypothesis in the K-S test), and relative efficiencies are discussed. A selection is then made for use in the K-S test.

Two estimates of the variance can be calculated directly from the output $\dot{r}$. These are the sample variance S_r^2 and the zero-mean sample variance $\hat{\sigma}^2$. The respective equations are given by

$$S_r^2 = \frac{1}{N-(n_o+1)} \sum_{i=n_o}^{N-1} (r_i - \bar{r})^2 \qquad (65)$$

and

$$\hat{\sigma}^2 = \frac{1}{N-n_o} \sum_{i=n_o}^{N-1} (r_i)^2 \qquad (66)$$

The only difference between the two equations is that S_r^2 is calculated around the sample mean,

$$\bar{r} = \frac{1}{N-n_o} \sum_{i=n_o}^{N-1} r_i \qquad (67)$$

while $\hat{\sigma}^2$ assumes the mean is zero.

It is a known fact (Ref. 7, pages 315-316) that

$$\frac{[N - (n_o+1)]S_r^2}{\sigma^2}$$

has a chi-square distribution with $N - (n_o+1)$ degrees of freedom while

$$\frac{(N - n_o)\hat{\sigma}^2}{\sigma^2}$$

has a chi-square distribution with $N-n_o$ degrees of freedom. (One degree of freedom is lost by S_r^2 because of the use of $\bar{r}$ in its calculation.)

Now the chi-square distribution with K degrees of freedom nas an expected value of K and a variance equal to 2K . Hence the expe ted value E and variance of $\tilde{\sigma}^2$ can be derived as follows:

$$\frac{(N-n_o)}{\sigma^2} E\tilde{\sigma}^2 = E\left(\frac{(N-n_o)\tilde{\sigma}^2}{\sigma^2}\right) = N-n_o$$

and

$$\frac{(N-n_o)^2}{\sigma^4} \text{VAR } \tilde{\sigma}^2 = \text{VAR}\left(\frac{(N-n_o)\tilde{\sigma}^2}{\sigma^2}\right) = 2(N-n_o)$$

Therefore

$$\left.\begin{aligned} E\tilde{\sigma}^2 &= \sigma^2 \\ \text{and} \qquad \text{VAR } \tilde{\sigma}^2 &= \frac{2\sigma^4}{N-n_o} \end{aligned}\right\} \qquad (68)$$

Similarly it can be shown that

$$\left.\begin{aligned} ES_r^2 &= \sigma^2 \\ \text{and} \qquad \text{VAR } S_r^2 &= \frac{2\sigma^4}{N-(n_o+1)} \end{aligned}\right\} \qquad (69)$$

The final candidate for a variance estimator is $\hat{\sigma}^2$, the square of the standard error of the estimate obtained from a regression analysis on the entire data base (i.e., sample size N). The equation for $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{1}{N-(P+1)} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \qquad (70)$$

where y_i is the actual cost of the ith procurement

$\hat{y}_i$ is the estimated cost of the ith procurement (obtained from a regression analysis of the entire sample)

and P is the number of independent variables in the linear model (PER)

It is known under the regression model assumptions (Ref. 7, page 364) that

$$\frac{[N - (P+1)]\hat{\sigma}^2}{\sigma^2}$$

has a chi-square distribution with $N - (P+1)$ degrees of freedom. Following the derivations of Eqs. 68 and 69 we then have

$$\left.\begin{aligned} E\hat{\sigma}^2 &= \sigma^2 \\ \mathrm{VAR}\ \hat{\sigma}^2 &= \frac{2\sigma^4}{N - (P+1)} \end{aligned}\right\} \qquad (71)$$

The facts obtained to this point are summarized in Table 26. We now address the question of which of these estimators is best to use in the K-S test.

As can be seen from the table, the candidate estimators are all unbiased,i.e., their expected value is σ^2 , the quantity that is being estimated. In addition, they are all consistent since the variance converges to zero as the sample size N gets large.

Difference in the estimators can, however, be seen when their relative efficiencies are examined. According to Ref. 7, page 216, if the estimators are unbiased, then the one with the smallest variance is

TABLE 26

CANDIDATE VARIANCE ESTIMATORS

Notation	Description	Equation	Expected Value	Variance
S_r^2	Sample variance of $\vec{r}$	64	σ^2	$\frac{2\sigma^4}{N - (n_o+1)}$
$\tilde{\sigma}^2$	Zero mean sample variance of $\vec{r}$	65	σ^2	$\frac{2\sigma^4}{N-n_o}$
$\hat{\sigma}^2$	Square of standard error of the estimate obtained from regression analysis of entire sample	69	σ^2	$\frac{2\sigma^4}{N - (P+1)}$

more efficient. From Sec. III C, it was pointed out that the minimum sample size for Historical Simulation must be larger than the number of parameters to be estimated. In the case being considered, P+1 parameters are to be estimated, one for each independent variable and one for the constant term. Hence $n_o > P+1$. This implies that for any sample size N , VAR $\hat{\sigma}^2$ < VAR $\tilde{\sigma}^2$ < VAR S_r^2 . Hence, $\hat{\sigma}^2$ is the most efficient of the three candidates.

Using efficiency as the criterion, $\hat{\sigma}^2$ would then be selected as the estimator of the variance σ^2 . It will be noted, however, that as N gets large, the differences in the variance of the estimators gets small, for example

$$\frac{\text{VAR } \hat{\sigma}^2}{\text{VAR } \tilde{\sigma}^2} = \frac{N - (P+1)}{N-n_o}$$

converges to 1 as N gets large. Hence, the advantage in efficiency for $\hat{\sigma}^2$ is only significant for small N . This is of course precisely the situation usually faced by the cost analyst.

Another advantage for choosing $\hat{\sigma}^2$ over $\tilde{\sigma}^2$ or S_r^2 is that $\hat{\sigma}^2$ is a function of the fit residuals from a regression analysis on the entire sample, rather than the prediction residuals from Historical Simulation. Hence it is more independent (in the non-statistical sense) of the prediction residuals than the other estimators , since $\hat{\sigma}^2$ does not depend directly on the values in the Historical Simulation residual vector $\vec{r}$. Therefore, $\hat{\sigma}^2$ more closely represents the given value (as compared to an estimate) of σ^2 that is called for in the K-S test.

For the reasons discussed above, $\hat{\sigma}^2$ has been selected as the estimate of σ^2 for the K-S test.

REFERENCES

1. C. A. Graver, Progress Report on the Development of Historical Simulation, General Research Corporation IMR-950, March 1969.

2. W. J. Dixon (Ed.), BMD Biomedical Computer Programs, University of California Press, 1968.

3. E. M. L. Beale, M. G. Kendall, D. W. Mann, The Discarding of Variables in Multivariate Analysis, Biometrika 54 (1967) 357-366.

4. C. A. Graver, The Use of Classical Statistics in Deriving and Evaluating CERs, Department of Defense (OASD) TP 66-8, October 1966.

5. C. A. Graver and H. E. Boren, Jr., Multivariate Logarithmic and Exponential Regression Models, The RAND Corporation, RM-4879-PR, July 1967.

6. N. R. Draper and H. Smith, Applied Regression Analysis, John Wiley & Sons, Inc., New York, 1968.

7. B. W. Lindgren, Statistical Theory, The MacMillian Company, New York, 1962

8. H. Cramér, Mathematical Methods of Statistics, Princeton University Press, 1946.

9. S. Siegel, Nonparameteric Statistics for the Behavior Sciences, McGraw-Hill, 1956.

10. D. A. Darling, "The Cramer-Smirnov Test in the Parametric Case," Annals of Mathematical Statistics, 26 (1955) Pg. 1.

11. J. M. McCormick and M. Salvadori, Numerical Methods in Fortran, Prentice Hall, 1964.

12. T. W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., March 1966.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
General Research Corporation
P.O. Box 3587, Santa Barbara, California 93105

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
-

3 REPORT TITLE
Historical Simulation: A Procedure for the Evaluation of Estimating Procedures
Vol. I of II - Procedure Development and Description

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final Report

5 AUTHOR(S) (First name, middle initial, last name)
C. A. Graver

6 REPORT DATE
June 1969

7a. TOTAL NO. OF PAGES
150

7b. NO. OF REFS
12

8a. CONTRACT OR GRANT NO.
DAHC15-68-C-0364

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
CR 0364-1, Vol. I of II

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 DISTRIBUTION STATEMENT

11 SUPPLEMENTARY NOTES

12 SPONSORING MILITARY ACTIVITY
Office of Assistant Secretary of Defense
Systems Analysis
The Pentagon, Washington, D.C.

13 ABSTRACT

A recurring problem faced by many analysts is that of devising estimating procedures for predicting some aspect of the future from rather meager data. This is particularly true for the cost analyst who is concerned with estimating the resource requirements of future military systems.

Historical Simulation is a method of evaluating candidate (cost) estimating procedures on the basis of their ability to simulate predictions using data that would have been available. In this fashion, Historical Simulation avoids relying on the central evaluation assumption of Regression Theory, namely, that which fits the past data best will predict the future best. This conceptual difference gives Histroical Simulation several unique features.

The report is divided into two volumes. The first volume, which is unclassified, completely describes the technique and unique features. Volume two, classified confidential (privileged information), illustrates the use of Historical Simulation by describing the results of applying the technique to cost and man-hour estimating procedures for selected aircraft programs. A non-technical overview of the contents of these documents can be found in General Research Corporation IMR-950.

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Prediction						
Regression						
Stepwise Multiple Regression						
Statistics						
Estimating						
Cost Estimating Re'ationship (CER)						